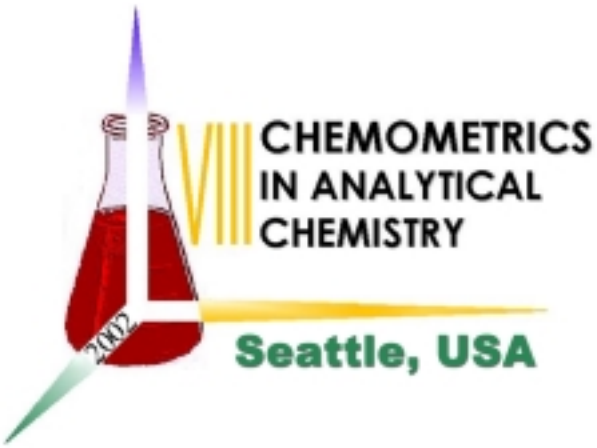

CAC-2002

Eighth International Conference on Chemometrics in Analytical Chemistry

Seattle, Washington, USA

September 22-26, 2002

Organized under the auspices of the North American/International Chemometrics Society (NAmICS)

Table of Contents

Sunday, September 23

17:00-19:00 Registration and Welcoming Reception

Monday, September 23

8:30-8:45 Welcome Address
Barry M. Wise

Session 1 ***Image Analysis***
Phillip K. Hopke, Chair

8:45-9:25 Exploration, Visualization and Resolution of Spectroscopic Images--Anna de Juan

9:25-9:50 Multivariate Curve Resolution for Hyperspectral Image Analysis: Applications to Microarray Scanner Fluorescence Images--David M. Haaland

9:50-10:15 Estimation of Trace Vapor Analyte Concentration-Pathlength for Remote Sensing Applications from Hyperspectral Images--Neal B. Gallagher

10:15-10:45 **Coffee Break**

10:45-11:10 Varimax Extended Rotation Applied to Multivariate Spectroscopic Image Analysis--Barry K. Lavine

11:10-11:35 Hyperspectral Imaging in a Forensic Laboratory: Data Cube Analysis Methods--Thomas W. Brueggemeyer

11:35-12:00 Self-Modeling Image Analysis with SIMPLISMA--Willem Windig

12:00-13:00 **Lunch**

13:00-14:00 **Poster Session**

Session 2 ***Process Chemometrics***
Paul J. Gemperline, Chair

14:00-14:40 Finding a Needle in a Haystack (A Challenge for Process Chemometrics)--Elaine B. Martin

14:40-15:05 The Acid Test for a Chemometric Method: Monitoring HF Alkylation Fluid via NIR Spectroscopy--Alan D. Eastman

15:05-15:30 Inferential Sensors Based on Integrating Analytical Neural Networks, Genetic Programming, and Support Vector Machines--Alex N. Kalos

15:30-16:00 **Coffee Break**

16:00-16:25 An Integrated Chemometrics Approach to Process Development for the Industrial Production of Antibiotics--Jose C. Menezes

16:25-16:50 Dynamic Time Warping of Spectroscopic Data for Statistical Batch Process Monitoring--Henk-Jan Ramaker

16:50-17:15 Detection and Correction of Non-calibrated Spectral Features in On-line Processes Spectroscopy--Frank Vogt

17:15-19:00 **Poster Session Mixer**

Tuesday, September 24

Session 3	***Multi-way and Curve Resolution*** Sarah Rutan, Chair
8:30-9:10	Estimation of Error Propagation and Prediction Intervals in Multivariate Curve Resolution Alternating Least Squares Using Resampling Methods--Rom Tauler
9:10-9:35	Advances in Hard- and Soft-Modeling of Multivariate Data--Marcel Maeder
9:35-10:00	Integrating Chemometrics with Chemical Separation Techniques--Robert E. Synovec
10:00-10:25	PARAFAC and Missing Values--Giorgio Tomasi
10:25-10:55	**Coffee Break**
10:55-11:20	Maximum Likelihood Parallel Factor Analysis (MLPARAFAC)--Lorenzo J. Vega-Montoto
11:20-11:45	The Influence of Data Fusion on Multi-Way Analysis of LC-DAD-MS Data--Ernst Bezemer
11:45-13:00	**Poster Session**
13:00-14:00	**Lunch**
Session 4	***Robust and Graphical Methods*** David L. Duewer, Chair
14:00-14:40	Data Visualization via Sufficient Dimension Reduction--Dennis Cook
14:40-15:05	The Use of Parallel Coordinate Graphical Plotting Combined with Principal Component Analysis Sample Scores for Visualizing Your Data at a Single Glance--Anthony D. Walmsley
15:05-15:30	A Comparative Study of Robust Estimation Methods--Anita Singh
15:30-16:00	**Coffee Break**
16:00-4:25	Exploring Data Set Structure with Density-based Approaches--Michal Daszykowski
16:25-16:50	Independent Component Analysis and Regression: Applications in Analytical Chemistry--Frank Westad
16:50-17:15	Exploratory Data Analysis of Spectra-structure Similarities--Kurt Varmuza
19:00-22:00	**CAC Night at Seattle Mariners**

Wednesday, September 25

Session 5	***Environmental Applications*** Cliff Spiegelman, Chair
8:30-9:10	Utilizing Hourly Gaseous Measurement as an Independent Variable in Multilinear Receptor Model Studies--Philip K. Hopke
9:10-9:35	A N-Way Analysis Technique of Two Tensors Applied to Ozone Concentration Analysis in the Paris Area--Georges Oppenheim
9:35-10:00	Comparison of Factor Analysis Methods for Evaluating the Trace Element Chemistry of Groundwaters of Southern Nevada-- Irene M. Farnham
10:00-10:25	Subsampling Particulate Samples: Theoretical Approximations for Environmental Matrices--Robert W. Gerlach
10:25-10:55	**Coffee Break**
Session 6	***Bioanalytical*** Peter de B. Harrington, Chair
10:55-11:35	Chemometric Opportunities in Proteomics--Alfred L. Yergey
11:35-12:00	Validation of Consensus Between Proteomic Expression and Clinical Chemical Data by a New Randomization F-test in Generalized Procrustes Analysis--Wen Wu
12:00-12:25	Use of Kinetic Equations in Analytical Clinical Chemistry-- Jeffrey E. Vaks
12:25-12:50	Modeling the Dynamic Effect of Tea in the Human Body Using Metabonomics - An Exploratory Study--Lefteris Kaskavelis
12:50-14:00	**Lunch**
Session 7	***Genetic Algorithms, Neural Networks and Datamining*** Steven D. Brown, Chair
14:00-14:40	Support Vector Machines for the Classification of Electronic Nose Data--Matteo Pardo
14:40-15:05	Selection of the Optimal Inputs in Chemometrics Modeling by Artificial Neural Network Analysis--Zvi Boger
15:05-15:30	Real-time Chemometrics Applied for Screening Food-Borne Pathogens and Bacterial Biomarker Using Ion Mobility and Differential Mobility Spectrometries: Chemometrics^{N}--Peter de B. Harrington
15:30-16:00	**Coffee Break**
16:00-16:25	Growing Neural Networks for Feature Selection and Calibration of Sensor Set-ups--Frank Dieterle
16:25-16:50	The Use of Continuous and Discrete Variables for Regression and Classification in Bayesian Networks--Nathaniel A. Woody
16:50-17:15	The O-PLS Approach, a New Modeling Concept in Multivariate Calibration--Johan Trygg
18:00-22:00	**Conference Dinner at Odyssey, the Maritime Discovery Center**

Thursday, September 26

Session 8	***Calibration*** Anthony D. Walmsley Chair
8:30-9:10	Wavelet Multiscale Regression Analysis for Multivariate Calibration--Steven D. Brown
9:10-9:35	Multivariate Calibration With Incomplete Designs--Clifford H. Spiegelman
9:35-10:00	Fast Algorithm for the Solution of Inequality Constrained Least Squares Problems--Mark Van Benthem
10:00-10:25	Transformation of Sensor Array Pattern Vectors into Descriptors of Unknown Vapors using Classical Least Squares and Inverse Least Squares Methods--Jay W. Grate
10:25-10:55	**Coffee Break**
10:55-11:20	Importance of Spectral Errors on Predictions Obtained by Partial Least Squares Regression--Charlotte M¿ller Andersen
11:20-11:45	Improving Piecewise Orthogonal Signal Correction--Huwei Tan
11:45-12:10	Calibration and Instrumental Design Strategies for Physiological Glucose Measurements Based on Near-IR Spectroscopy--Gary W. Small
12:10-12:15	**Official Closing of CAC-2002**
12:15-13:15	**Lunch**

CAC-2002 Committees

CAC Permanent Committee

L. Buydens (NL)
M. Forina (I)
P. Hopke (USA)
D.L. Massart (B)
X. Rius (E)
H. Smit (NL)
B.G.M. Vandeginste (NL)
P. Van Espen (B)

International Scientific Committee

R. Brereton (GB)
J. Gasteiger (D)
J. Havel (CZ)
B. Kowalski (USA)
N. Maeder (AUS)
L. Munck (DK)
S. Rutan (USA)
R. Tauler (E)
K. Varmuza (A)
P.D. Wentzell (Canada)
J. Zupan (SLO)
S. Clementi (I)
P. Geladi (S)
O.M. Kvalheim (N)
Y. Liang (P.R. China)
P. Minkkinen (FIN)
R. Phan Tan Luu (F)
A. Smilde (NL)
J. Tetteh (GB)
B. Walczak (PL)
S. Wold (S)

Local Organising Committee

Barry M. Wise, chair (USA)
Neal B. Gallagher (USA)
Mary Beth Seasholtz (USA)
Peter de B. Harrington (USA)
Karl Booksh (USA)
Nan Holmes (USA)
Terry Lane, secretary (USA)
Mel Koch (USA)
David L. Duewer (USA)
Paul J. Gemperline (USA)
David Veltkamp (USA)
James Jordan (USA)

CAC-2002 Sponsors

The organizers of CAC-2002 would like to thank our generous sponsors for their support. Contributed funds were used to help defray the travel expenses of invited speakers and for group events such as the conference dinner. CAC-2002 is supported by Underwriters (contributions of $5K or more), Sponsors (contributions of $2K - $ 5K) and Contributors (contributions < $2K). Our sponsors include:

CAC-2002 Underwriters:

Dow

U.S. Environmental Protection Agency

Unilever

CAC-2002 Sponsors:

Elsevier Science

Eigenvector Research, Inc.

CAC-2002 Sponsors (continued):

ExxonMobil

Pfizer

CAC-2002 Contributors:

Infometrix

Royal Society of Chemistry

Measurement & Control Engineering Center

DuPont

Center for Process Analytical Chemistry

The MathWorks, Inc.

Oral Program Abstracts

Monday, September 23

8:30-8:45 Welcome Address
Barry M. Wise

Session 1 *Image Analysis*
Phillip K. Hopke, Chair

8:45-9:25 Exploration, Visualization and Resolution of Spectroscopic Images
Anna de Juan

9:25-9:50 Multivariate Curve Resolution for Hyperspectral Image Analysis: Applications to Microarray Scanner Fluorescence Images
David M. Haaland

9:50-10:15 Estimation of Trace Vapor Analyte Concentration-Pathlength for Remote Sensing Applications from Hyperspectral Images
Neal B. Gallagher

10:15-10:45 Coffee Break

10:45-11:10 Varimax Extended Rotation Applied to Multivariate Spectroscopic Image Analysis
Barry K. Lavine

11:10-11:35 Hyperspectral Imaging in a Forensic Laboratory: Data Cube Analysis Methods
Thomas W. Brueggemeyer

11:35-12:00 Self-Modeling Image Analysis with SIMPLISMA
Willem Windig

12:00-13:00 Lunch

13:00-14:00 Poster Session

Title: Exploration, visualization and resolution of spectroscopic images

Authors: Anna de Juan[1], Marcel Maeder[2], Rom Tauler[1]

1 Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

2 Department of Chemistry; University of Newcastle; Callaghan, NSW 2308; Australia

Keywords: image analysis, EFA, curve resolution, two-way, three-way

Presenter: Anna de Juan, annaj@apolo.qui.ub.es

Spectroscopic imaging is increasingly used in most varied fields to describe globally and locally samples of all sizes and origins. A simple image can easily consist of thousands of pixels with their related spectra, *i.e.* a massive experimental output with no straightforward useful information. Chemometrics applied to image analysis should address two main issues: the sensible size reduction of the original image without losing relevant information and the adaptation or development of data analysis tools that take into account the particular features of image data sets.

Data size reduction is preferably performed on the spectral direction to avoid losses in the spatial resolution, most important in image description. Compressed multivariate spectral representations obtained by wavelet transformation or Principal Component Analysis are proposed to replace rough univariate representations of the pixels using supposedly selective wavelengths or cumulative intensities.

Surface and multilayer images are graphically displayed as cubes and hypercubes, respectively, whose dimensions are the wavelengths and the two or the three spatial pixel coordinates. However, surface images form actually two-way data sets that follow a bilinear model, where the mixed signal recorded in each pixel is described by the concentration-weighted sum of the pure signals of the chemical compounds present. Analogously, multilayer images organize as three-way data sets, whose third direction accounts for layer-to-layer changes in the bulk abundance of each image compound.

Adaptations of exploratory tools based on global and local rank analysis and of resolution methods are described using real examples of two- and three-way images of pharmaceutical, industrial and biomedical origin. The chemometric tools proposed work with methods based on bilinear models and extensions of those to deal with three-way data sets. However, the way to organize and handle the information is devised so as the spatial structure of the image plays an active role in the information obtained and the visualization of the results is also designed to match the spatial image structure.

Title: Multivariate curve resolution for hyperspectral image analysis: Applications to microarray scanner fluorescence images

Authors: David M. Haaland[1], Jerilyn A. Timlin[1], Michael B. Sinclair[1], Mark H. Van Benthem[1], Juanita Martinez[2], Angelina Rodriguez[2], Anthony Aragon[2], Angelina Rodriguez[2], Jose˚Weber[2], Margaret Werner-Washburne[2]

1 Sandia National Laboratories; Mail Stop 0886; Albuquerque, NM 87185-0886; USA

2 University of New Mexico, Albuquerque, NM 87131

Keywords: image analysis, curve resolution

Presenter: David Haaland, dmhaala@sandia.gov

Multivariate curve resolution (MCR) using constrained alternating least squares algorithms represents a powerful analysis capability for the quantitative analysis of hyperspectral image data. We will demonstrate the application of MCR using data from a new hyperspectral fluorescence imaging microarray scanner for monitoring gene expression in cells from thousands of genes on the array. We will present the design of the new scanner, that collects the entire fluorescence spectrum from each pixel of the scanned microarray. The use of MCR with non-negativity and equality constraints has allowed us to detect contaminating fluorescence that emits at the same wavelengths as the reporter fluorophores on microarray slides. We will show that the contaminant fluorescence under the printed DNA spots is not the same as the region next to the spots. Thus, traditional background subtraction methods used with data collected from the current commercial microarray scanners are often not correct and will lead to errors in determining the relative expression of low-expressed genes.

With the new scanner, we are able to generate relative concentration maps of the background, impurity, and fluorescent labels over the entire image. Since the concentration maps of the fluorescent labels are relatively unaffected by the presence of background and impurity emissions, the accuracy and useful dynamic range of the gene expression data are both greatly improved over those obtained by commercial microarray scanners that produce univariate images for each fluorescent label. MCR methods to achieve quantitative results over a wide range of microarray data will be presented. Enhancements to the MCR algorithms used in this work include improvements in computational efficiency, use of equality constraints with baseline variations and known emissions, proper weighting of the emission data, and methods designed to greatly improve the dynamic range of the instrument.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000. This project also supported in part by a grant from the Keck Foundation.

Title: Estimation of trace vapor analyte concentration-pathlength for remote sensing applications from hyperspectral images

Authors: Neal B. Gallagher[1], David M. Sheen[2]

1 Eigenvector Research, Inc.; PO Box 561; Manson, WA 98831; USA

2 Battelle Pacific Northwest National Laboratory; PO Box 999, Mail Stop K5-25; Richland, WA 99352; USA

Keywords: remote sensing, image analysis, hyperspectral images, augmented least squares, error propagation

Presenter: Neal B. Gallagher, nealg@eigenvector.com

Near infrared hyperspectral imaging is finding utility in remote sensing applications such as detection and quantification of chemical vapor effluents in stack plumes. An algorithm based on a novel application of augmented least squares for estimating the on-plume pixel background clutter and concentration-pathlengh is presented. The model is based on radiance equations and Beer s Law at low concentration-pathlength. This algorithm requires estimates of the atmospheric radiance and transmission from the target plume to the imaging spectrometer, and an estimate of the plume temperature. However, a distinct advantage is that it does not require estimates of the background temperature or emissivity. The algorithm effectively provides a local estimate of the clutter and an error analysis shows that it can provide superior quantification over approaches that model the background clutter in a more global sense.

The error analysis used a detailed noise model based on Poisson statistics for both a passive Fourier transform spectrometer and a dispersive spectrometer, and synthetically generated hyperspectral images. Error analysis and tests using synthetic hyperspectral images were used to provide a comparison of estimation algorithms. These tests showed that the proposed algorithm provided good estimates of concentration-pathlength and was significantly better than global approaches.

Title: Varimax extended rotation applied to multivariate spectroscopic image analysis

Authors: Barry K. Lavine, Charles E. Davidson, Jason Ritter
Department of Chemistry; Clarkson University; Potsdam, NY 13699-5810; USA

Keywords: MCR, varimax rotation, extended rotation, end members, imaging, diode array data

Presenter: Barry Lavine, bklab@clarkson.edu

A robust approach to spectral image analysis called the Varimax extended rotation (VER) method has been developed. VER utilizes a four-step procedure to resolve 2-way data. In the first step, the data is preprocessed to ensure that it is in a form suitable for multivariate curve resolution. The second step involved principal component analysis, reducing the dimensionality of the data while retaining the information in the data. In the third step, a new coordinate system is developed to display the data using a Varimax extended rotation that assists in the identification of the so-called pure regions in the data matrix while simultaneously rotating both the concentration and spectral matrices towards a solution. The fourth and final step involves the use of alternating least squares (ALS) algorithm that improves the estimates of both the concentration and spectral profiles of each component. Using HPLC diode array data of multicomponent mixtures and Raman imaging data of water in oil emulsions, we will demonstrate the efficacy and advantages of the VER approach.

Title: Hyperspectral imaging in a forensic laboratory: Data cube analysis methods

Authors: Thomas W. Brueggemeyer, Mark R. Witkowski, Jonathan Litzau
U.S. Food & Drug Administration; 6751 Steger Drive; Cincinnati, OH 45237; USA

Keywords: imaging, hyperspectral, correlation

Presenter: Thomas Brueggemeyer, tbruegge@ora.fda.gov

Hyperspectral imaging with chemometric data analysis promises to play an important role in the forensic laboratory. The software provided with commercial spectrometers may be less than ideal for the applications discussed below. Data cube analysis software implemented in the Matlab language offers the flexibility required to consider both new algorithms and modifications to existing ones.

One important application of hyperspectral imaging involves finding and then identifying isolated contaminants in an otherwise normal sample matrix. The techniques frequently utilized in such imaging — FTIR, Near-IR, and Raman spectroscopies — are not known for extreme sensitivity. However, an image in which one or more pixels are spectrally distinct from the others can be indicative of contamination, adulteration, tampering, or counterfeiting. Thus, a component present at a bulk concentration inadequate for traditional spectroscopic detection, may nevertheless be seen in an image if it is heterogeneously distributed.

From a chemometrics perspective, contrast maximization is the goal when discrete particles or regions are being sought in an image. However, the approach to be undertaken will vary depending upon which of two scenarios is in effect. In the first scenario, there is a specific, targeted analyte whose spectrum has been established. In this case, various types of correlation mapping can be employed, indicating the degree of similarity to the target spectrum for each pixel in the image. Thus the data cube is compressed to a single-plane similarity image. Different similarity measures will be discussed.

In a second scenario, images of interest are examined for the presence of unknown particles that stand out from the sample matrix as a whole. Here, without a target spectrum, differences from the bulk sample become the focus. Hence, data cube compression is based upon a model of the overall spectral variability in the image. Pixels well-removed from the model space are suspect. The utility of different algorithms for best highlighting unknown contaminant particles will be discussed.

Title: Self-Modeling Image Analysis with SIMPLISMA

Authors: W. Windig[1], S. Markel[2], P.M. Thompson[2]

1 Eigenvector Research, Inc., East Coast Office; 6 Olympia Drive; Rochester NY 14615; USA

2 Imaging Materials and Media; Research & Development; Eastman Kodak Company; Rochester, NY 14650-2132; USA

Keywords: self-modeling image analysis (SIA), self-modeling mixture analysis (SMA), SIMPLISMA, FTIR, SIMS

Presenter: Willem Windig, windig@eigenvector.com

Multivariate image analysis can be done with techniques such as principal component analysis (PCA) [1]. The major advantage to this approach is that many images (one image per wavelength) can be reduced to a few images, while preserving all significant information in the data set. Furthermore, noise reduction is obtained. The data reduction obtained with PCA, however, is an orthogonal system that explains the maximum variance only. Therefore, the reduced numbers of images are difficult to interpret.

Another approach is to use self-modeling mixture analysis (SMA). In this case, the data set is reduced to a limited number of images, based on mathematical criteria that produce chemically meaningful results. For example, a chemical image of a cross section of three different polymers results in three images, each describing one polymer. Each of the images has an associated spectrum describing the pure polymer. Examples of self-modeling image analysis (SIA) have been described by Sasaki et al. [2], where entropy minimization was used to resolve 13 images (400-700 nm at 25-nm intervals) of biological samples. Andrew and Hancewicz [3] analyzed Raman Image Data by selecting the most dissimilar spectra in a data set. Batonneau et al. selected pure variables from Raman spectroscopic images of industrial dust particles [4] using SIMPLISMA (simple-to-use interactive self-modeling mixture analysis). The latter technique will is also used for this study. SIMPLISMA is a technique that selects pure variables from the spectral data set. Pure variable values are proportional to concentrations and can be used to resolve data. When no pure variables are present, second derivative data can be used. SIMPLISMA has been used to image related projects: resolving FTIR (Fourier transform infrared) microscopy data of a polymer laminate [5,6], and to resolve FTIR microscopy data of a KBr tables with a mixture of three powders [6].

A newly developed SIMPLISMA addition is used in this study [7]. The new addition allows the use of a combination of conventional and second derivative data, enabling the user to extract baseline-related problems. Examples are given of:

a) FTIR microscopy transmission data of a cross section of a polymer laminate,
b) FTIR microscopy reflectance data of a mixture of aspirin and sugar, and
c) SIMS (secondary ion mass spectrometry) of a two-component mixture of palmitic and stearic acids.

In the latter case, the differences are shown between selecting pure spectra and pure variables (masses) as a starting point for resolving and selecting. Furthermore, it is shown how to obtain a pure variable solution with the pure spectra as a starting point.

1) P. Geladi, H. Grahn. Multivariate Image Analysis. John Wiley & Sons, 1996.
2) K. Sasaki, S. Kawata, S. Minami. J Opt Soc Am A 1989;6:73-79.
3) J.J. Andrew and T.M. Hancewicz. Appl Spectrosc 1998;52:797-807.
4) Y. Batonneau, J. Laureyns, J.C. Merlin, C. Bremard. Anal Chim Acta 2001;446:23-37.
5) W. Windig, S. Markel. J Mol Struct 1993;292:161-170.
6) J. Guilment, S. Markel, W. Windig. Appl Spectrosc 1994;48:320-326.
7) W. Windig, B. Antalek, J.L. Lippert, Y. Batonneau, C. Bremard. Anal Chem 2002;74:1371-1379.

Monday, September 23

Session 2 *Process Chemometrics*
Paul J. Gemperline, Chair

14:00-14:40 Finding a Needle in a Haystack (A Challenge for Process Chemometrics)
Elaine B. Martin

14:40-15:05 The Acid Test for a Chemometric Method: Monitoring HF Alkylation Fluid via NIR Spectroscopy
Alan D. Eastman

15:05-15:30 Inferential Sensors Based on Integrating Analytical Neural Networks, Genetic Programming, and Support Vector Machines
Alex N. Kalos

15:30-16:00 Coffee Break

16:00-16:25 An Integrated Chemometrics Approach to Process Development for the Industrial Production of Antibiotics
Jose C. Menezes

16:25-16:50 Dynamic Time Warping of Spectroscopic Data for Statistical Batch Process Monitoring
Henk-Jan Ramaker

16:50-17:15 Detection and Correction of Non-calibrated Spectral Features in On-line Processes Spectroscopy
Frank Vogt

17:15-19:00 Poster Session Mixer

Title: Finding a needle in a haystack (A challenge for process chemometrics)

Authors: Elaine B. Martin, Julian Morris
CPACT; University of Newcastle; Merz Court; Newcastle upon Tyne NE1 7RU; United Kingdom

Keywords: process monitoring, super model-based PCA (SMBPCA), fermentation

Presenter: Elaine B. Martin, e.b.martin@ncl.ac.uk

The widespread use of batch processes in the process manufacturing industries for the production of speciality products has led to an increase in the need for the application of theoretically valid process performance monitoring tools in order to contribute to the achievement of consistent, high quality production. The linear multivariate statistical batch monitoring techniques of multi-way principal component analysis (MPCA) and multi-way partial least squares (MPLS) (Nomikos and MacGregor, 1994, 1995) have been the standard approaches used by both by academic researchers and industry for the modeling and monitoring of batch processes. However, in practice data collected on batch processes typically exhibit non-linear, time-variant behavior. These properties challenge the traditional multi-way techniques for the assured modeling and monitoring of batch processes. The presentation will demonstrate the limitation of the existing approaches through a benchmark simulation study and an application to an industrial fed-batch process.

Two solutions to the problem will then be presented, super model- and local model-based process performance monitoring. In model-based performance monitoring a reduced complexity phenomenological model is fitted to the process data giving a set of residuals that can be used as the basis for a monitoring scheme. To allow for the modeling errors, this approach has been extended to a novel technique termed Super Model-Based PCA (SMBPCA). SMBPCA includes an additional residuals modeling stage, allowing an imperfect, reduced complexity mechanistic model to be used in the monitoring scheme. SMBPCA shows enhanced process performance monitoring capabilities in comparison to conventional multivariate performance monitoring techniques with respect to fault detection and false alarm rate. The methodology is demonstrated on a benchmark simulation of a batch reactor.

An alternative approach to modeling dynamic non-linear batch systems is to sub-divide the process variable trajectories into a number of distinct operating regions. In this situation a local linear model can then be fitted to each separate region. These individual models can then be pieced together, thereby providing an overall non-linear global model (e.g. Foss et al., 1995). This piecing together of the local models can be achieved through a number of means including validity functions and fuzzy membership functions (e.g. Murray-Smith and Johansen, 1997). These functions are then used to determine which local model is the most appropriate for representing a particular region in the batch trajectory. Such a local model based structure realizes a novel approach to dynamic batch process performance monitoring. An industrial fed-batch fermentation process is used to demonstrate the methodology.

Title: The acid test for a chemometric method: Monitoring HF alkylation fluid via NIR spectroscopy

Authors: Alan D. Eastman, Randall L. Heald, W. Pat Moore, Bruce B. Randolph
Phillips Petroleum Co.; 152 Petroleum Lab; Bartlesville, OK 74004; USA

Keywords: HF, ASO, water, alkylation, designed experiment

Presenter: Alan Eastman, adeastm@ppco.com

The concentrations of HF acid, acid-soluble oil (ASO), and water in a large commercial HF alkylation reactor have been determined using an on-line near-infrared analyzer. The partial least-squares model was constructed from a designed experiment utilizing gravimetric blends of the components, then modified by inclusion of one sample from the plant. After startup, it became clear that on-line analysis provides unit operators better data than that from the refinery laboratory. Instrument stability and maintenance problems in the plant setting are also discussed.

In essence, a successful chemometric model is only the starting point for a successful chemometric application in the real world. This system is now being marketed worldwide, with a number of operating units.

Title: Inferential sensors based on integrating analytical neural networks, genetic programming, and support vector machines

Authors: Alex N. Kalos, A.K. Kordon, G.F. Smits, S. Werkmeister
The Dow Chemical Company; 2301 N. Brazosport Boulevard; Bldg B1217;
Freeport, TX 77541; USA

Keywords: soft sensor, inferential sensor, stacked analytic NN, genetic programming, support vector machines

Presenter: Alex Kalos, ankalos@dow.com

Soft (or inferential) sensors assume that there is an empirical relationship between some easily measured and continuously available process variables and some critical parameters related to process or product quality. A novel methodology for development of inferential sensors through integration of Stacked Analytical Neural Networks (SANN), Support Vector Machines (SVM), and Genetic Programming (GP) is proposed.

The objective of the integration is to enhance the advantages of each approach, to reduce development time, and to increase the robustness of the on-line model. SVM contribute mainly to outlier detection and compression of the data set to the most informative data. SANN perform fast nonlinear sensitivity analysis and deliver a black box -type inferential sensor with model agreement confidence indicator. GP generates an inferential sensor, via symbolic regression, in explicit functional form that can be implemented directly in a Distributed Control System.

The main advantages of this type of inferential sensors are their good generalization capabilities, explicit input/output relationships, and low implementation and maintenance cost. GP-generated inferential sensors have the potential to be more robust for real industrial applications than traditional neural nets alone. Of particular significance is the ability to examine the behavior of the model outside the training range. This can be done easily and directly with a function-based solution, whereas it is difficult in the case of a black-box model.

An example of a robust inferential sensor for interface level estimation, developed by the proposed methodology in The Dow Chemical Company, is given. It is a novel approach to empirical model building as it is based on operators estimates for output values, rather than analytical measurements. The inferential sensor has self-assessment capability via a parameters-within-range indicator that protects process operators from unreliable predictions. Since its on-line installation, the inferential sensor shows robust performance during the production of several different product grades and transitions between them.

Title: An integrated chemometrics approach to process development for the industrial production of antibiotics

Authors: J.A. Lopes, Jose C. Menezes
Department of Chemical Engineering; Centre for Biological& Chemical Engineering; Instituto Superior Tecnico; Av. Rovisco Pais; P1049-001 Lisbon; Portugal

Keywords: fermentation, modelling, data mining, NNs, MSPC

Presenter: Jose Menezes, bsel@ist.utl.pt

The batch production process of an antibiotic-like molecule by fermentation in an industrial environment was analyzed using chemometrics methods applied to a number of on- and at-line measurements (e.g. NIR). The process was analyzed from a global perspective considering not only the antibiotic production stage but also previous stages. We were able to develop an integrated approach to perform tasks such as data-mining, process modeling and statistical process supervision.

Component models (principal components analysis, PARAFAC, Tucker3) were able to provide information about the process trends and potential for process optimization. The relative importance of each process stage and each raw material type on the process productivity was measured through the implementation of a multiblock partial least squares strategy. The main purpose of this model was to obtain estimates of the process productivity based on the expected feed profiles and raw materials quality.

Based on this information a set of experiments were designed to provide the data for process modeling. Based on these experiments neural networks were used to develop static and dynamic models for the fermentation process. Genetic algorithms were used to optimize the models inputs and the runs selected for calibration and validation. These models provided long term and one step ahead predictions of the main process state variables.

Multivariate statistical process supervision models based on a linear method (PARAFAC) and a non-linear component method (autoassociative neural network) were developed and applied to follow the process on-line in order to prevent process deviations.

We describe the advantages of combining several chemometrics tools and the principal benefits achieved with each strategy. We also depicted a general overview of the principal achievements and limitations of the techniques used within the presented methodology.

Title: Dynamic time warping of spectroscopic data for statistical batch process monitoring

Authors: Henk-Jan Ramaker, Eric van Sprang, Age Smilde
Department of Chemical Engineering; University of Amsterdam;
Nieuwe Achtergracht 166; NL1018 WV Amsterdam; The Netherlands

Keywords: dynamic time warping, statistical batch process monitoring, MSPC

Presenter: Henk-Jan Ramaker, hj.ramaker@science.uva.nl

Batch processes are very common in chemical, pharmaceutical, food industry and biochemistry. Monitoring these batch processes is wanted for several reasons such as safety, waste-stream reduction, consistency, quality improvement or improved process knowledge. One of the methods for batch process monitoring is based upon multivariate techniques and was introduced by Nomikos and McGregor (1994). This technique is referred to as statistical batch process monitoring. Many applications and extensions of statistical batch process monitoring have been introduced since.

Most common batch processes are equipped with sensors that measure engineering variables like e.g. temperatures, pressures and flow rates. However, modern process analyzers, like spectroscopic measurement devices, find their way into batch monitoring. The main advantage of these apparatus is that the measurements contain chemically richer information compared to engineering variables.

For statistical batch monitoring a database is needed with completed batch runs that produced on-spec products. The variation within this data serves as a reference distribution. The performance of independent new batches are compared with this reference distribution. This is achieved using multivariate control charts. These control charts are normally based on PCA or PLS models. For some models it is required that the reference batches have equal duration. However, in practice this is almost never true. Therefore the models for statistical batch monitoring cannot be used.

The problem of batches with unequal lengths can be overcome by the choice of model. However, the models capable of dealing with batches of unequal length suffer from poor statistics. Another way to deal with this issue is to make use of dynamic time warping (DTW). DTW originates from the world of speech recognition and Kassidas et al. (1998) proposed a method for a DTW algorithm that is focused on statistical batch monitoring. This method was based on batch data that contained engineering variables.

In this work, data from an industrial batch reactor that produces a resin is used. This batch reaction is spectroscopically monitored with NIR. The batches need to be warped because they have unequal length. A strategy is proposed how to warp spectroscopic data since this is not yet discussed in the research area of statistical batch process monitoring. Several constrains are applied and the quality of the DTW algorithm is expressed using performance indices.

Title: Detection and correction of non-calibrated spectral features in on-line processes spectroscopy

Authors: Frank Vogt, Boris Mizaikoff, Maurus Tacke
School of Chemistry and Biochemistry; Georgia Institute of Technology;
770 State Street; Atlanta, GA 30332-0400; USA

Keywords: non-calibrated spectral features, qualitative test of calibration model, wavelet analysis, process monitoring

Presenter: Frank Vogt, FRNVogt@aol.com

Spectroscopic monitoring techniques usually employ chemometric algorithms like principal component regression for calibration and evaluation of optical spectra. However, data evaluation quality of process measurement systems based on such algorithms is substantially affected by unknown spectral features appearing after calibration. Absorbers not contained in the calibration model may result in major deviations of concentration readings — on the other hand, their recognition may indicate problems, for instance related to chemical processes or during waste water analysis. Hence, detection of non-calibrated absorption features is of importance for enhanced quality control using spectroscopic monitoring techniques. A numerical evaluation technique augmented by a qualitative data analysis step provides additional information, enabling improved assessment of errors due to the occurrence of non-calibrated absorbers. Qualitative error analysis relies on gathering spectroscopic properties of the disturbance, *i.e.* wavelength position and shape of the non-calibrated features.

In principle, there are three constraints during the search for non-calibrated spectral features:

1) Data analysis must be independent from the concentration of the absorbers and from the composition of the samples.
2) On-line feasibility of such data analysis requires fully automated evaluation procedures.
3) The whole measured spectral range and all contained spectral features must be analyzed, ensuring that all non-calibrated features are discovered.

For this purpose a novel approach is proposed: Following conventional PCA, the wavelet representations of the resulting principal components simultaneously define the calibrated spectral features with respect to wavelength position and shape. By means of analyzing wavelet transformed unknown measurement spectra, it is shown that non-calibrated spectral features can be detected and characterized with respect to their wavelength position and shape.

Tuesday, September 24

Session 3 *Multi-way and Curve Resolution*
Sarah Rutan, Chair

8:30-9:10 Estimation of Error Propagation and Prediction Intervals in Multivariate Curve Resolution Alternating Least Squares Using Resampling Methods
Rom Tauler

9:10-9:35 Advances in Hard- and Soft-Modeling of Multivariate Data
Marcel Maeder

9:35-10:00 Integrating Chemometrics with Chemical Separation Techniques
Robert E. Synovec

10:00-10:25 PARAFAC and Missing Values
Giorgio Tomasi

10:25-10:55 Coffee Break

10:55-11:20 Maximum Likelihood Parallel Factor Analysis (MLPARAFAC)
Lorenzo J. Vega-Montoto

11:20-11:45 The Influence of Data Fusion on Multi-Way Analysis of LC-DAD-MS Data
Ernst Bezemer

11:45-13:00 Poster Session

13:00-14:00 Lunch

Title: Estimation of error propagation and prediction intervals in multivariate curve resolution alternating least squares using resampling methods

Authors: Joaquim Jaumot, Raimundo Gargallo, Rom Tauler
Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

Keywords: MCR-ALS, error propagation, resampling

Presenter: Roma Tauler, roma@apolo.qui.ub.es

Alternating Least Squares (ALS) has become a popular method for Multivariate Curve Resolution (MCR) mostly due to its flexibility in constraint implementation during the optimization of resolved profiles. A problem that remains unexplored using these techniques is the evaluation of the reliability of obtained estimations.

In this evaluation, two different but correlated aspects should be considered. The first aspect is the rotational and scale freedom of estimations, a common problem of factor analysis methods. The second aspect is how errors and noise are propagated from experimental data to ALS estimations. In this work, these two aspects are preliminarily investigated.

When noise levels in experimental data increase, both aspects are intermixed and it is rather difficult to discern between them and differentiate what causes what. Different approaches for calculation of error propagation and prediction intervals of estimations are explored and compared between them including, Monte Carlo simulations, resampling approaches and Jackknife based methods. The obtained results allowed a preliminary investigation of noise effects on resolved profiles and on parameters from them estimated, and allowed also a preliminary investigation of noise effects on rotational ambiguities. The study is shown for the resolution of a three-component equilibrium system with overlapping concentration and spectra profiles.

Title: Advances in hard- and soft-modeling of multivariate data

Authors: Marcel Maeder, Caroline Mason, Yorck-Michael Neuhold, Graeme Puxty, Raylene Dyson
Department of Chemistry; University of Newcastle; Callaghan, NSW 2308; Australia

Keywords: hard-modelling, soft-modelling, data fitting, global analysis, FA, buffer, kinetics, equilibria

Presenter: Marcel Maeder, chmm@cc.newcastle.edu.au

Data fitting (hard-modeling) has long been applied for the analysis of chemical data. The parameters of a given function are fitted in order to result in an optimal fit with the measured data. Many algorithms are readily available.

So-called chemometric methods are more recent developments; they are usually attempts to decompose the data into its relevant contributions without being based on a pre-selected function; the only restrictions are physically defined, such as positive concentrations and molar absorptivities.

Advances in hard-modeling comprise the global analysis of series of data sets. This can alleviate difficulties arising from linear dependencies or from badly defined minor species. Applications in reaction kinetics and equilibrium studies are presented. We will also present attempts to increase the complexity of the model, e.g. incorporating non-ideal behavior such as pH changes during the reaction. This allows the analysis of unbuffered kinetics, and thus avoids the necessity of the addition of buffers that often interfere with the reaction under investigation.

Soft-modeling methods are required if there are no functional relationships to describe the measurement. These methods suffer from a lack of robustness due to a very large numbers of parameters. This problem is addressed by the recent method of resolving factor analysis . Additionally, soft-modeling analyses are often not unique, as only bands of feasible solutions result. We will present novel ways of analyzing and representing these regions of possible solutions.

Title: Integrating chemometrics with chemical separation techniques

Authors: Robert E. Synovec, Bryan J. Prazen, Kevin J. Johnson, Bob W. Wright, Kristin˚H.˚Jarman
Department of Chemistry; University of Washington; Box 351700; Seattle, WA 98195; USA

Keywords: chromatography,tri-linear PLS, retention time alignment

Presenter: Robert Synovec, synovec@chem.washington.edu

Chemical separation techniques are ideally suited to the analysis of complex samples. The current mindset is to design separations that achieve baseline resolution of the analytes of interest from each other and from other components of the sample. While the current practice has been generally successful, full resolution of all of the analytes of interest in a complex real sample is often impractical or impossible, due to the excessively long analysis times required. In order to either reduce analysis time or simply to make feasible a particular analysis, it has become imperative to investigate novel problem-solving approaches.

Our goal is to combine separation methods with chemometrics, resulting in useful problem-solving strategies, with the potential to address these important chemical analysis issues. A major impediment to achieving this goal is run-to-run retention time variation. Comparison of different chromatograms with severe run-to-run retention time variation can significantly reduce the utility of chemometric algorithms. We have been studying the retention time reproducibility issue for both one-dimensional (e.g. gas chromatography, GC) and comprehensive two-dimensional separation techniques (e.g. GC x GC).

We have developed several different strategies to objectively correct for run-to-run retention time variation prior to applying chemometric methods. These strategies will be reported as they relate to concurrent developments in novel instrumentation. First, we report recent work in the area of retention time alignment implementing a rank minimization criteria. This alignment technique is coupled with the generalized rank annihilation method (GRAM) for the analysis of separations obtained by GC x GC. Second, we report the development of a strategy in which entire GC x GC chromatograms are aligned prior to calibration with tri-linear partial least squares (tri-PLS). This alignment strategy combines rank minimization with retention time axis stretching through interpolation. The third area we describe is a recently developed algorithm that objectively corrects very slight, yet significant, variations in run-to-run retention time for one-dimensional separations. The algorithm effectively preserves the desired sample-to-sample chemical selectivity while minimizing run-to-run retention time variation. Application of this one-dimensional alignment algorithm greatly enhances the quality of information obtained in pattern recognition studies utilizing principal components analysis (PCA). All of the algorithms are generally applicable to data collected from a variety of separation techniques, and are not necessarily limited to GC.

Title: PARAFAC and missing values

Authors: Giorgio Tomasi, Rasmus Bro
Department of Dairy and Food Science; Royal Veterinary and Agricultural University;
Rolighedsvej 30, iii; DK1958 Frederiksberg C; Denmark

Keywords: PARAFAC, missing values, expectation maximization, Levenberg-Marquadt, INDAFAC

Presenter: Giorgio Tomasi, gt@kvl.dk

Missing values are a common occurrence in chemometric data and different approaches have been proposed to deal with them in modeling. In this work two different ideas and two algorithms are compared in their efficiency in dealing with incomplete data. In the first, Expectation Maximization is combined with a standard PARAFAC-ALS algorithm whereas in the second a computationally more expensive method (*i.e.*, an appropriately modified Levenberg-Marquadt) fits the model only to the existing elements.

The performances of these two algorithms and the effect of the incompleteness of the data on the fitted models have been evaluated on the basis of both simulated and real data sets with different amounts (between 40% and 70%) and patterns (random or systematic) of missing values. The evaluation is based on the quality of the solution expressed in terms of accuracy of parameter estimates and fit. Some observations are also made in terms of computational efficiency of the two methods with respect to the number of iterations and the time consumption.

Title: Maximum likelihood parallel factor analysis (MLPARAFAC)

Authors: Lorenzo J. Vega-Montoto, Peter D. Wentzell
Department of Chemistry; Dalhousie University; Halifax, Nova Scotia B3H 4J3; Canada

Keywords: ALS, PARAFAC, MLPARAFAC

Presenter: Lorenzo J. Vega-Montoto, mvega@is2.dal.ca

In recent years, second and higher order instrumentation has become more commonplace in chemical research. Examples include fluorescence excitation-emission spectroscopy and chromatography instrumentation with multichannel detectors, used in a variety of arrangements to obtain second or higher order tensor data. The intrinsic multilinear structure of the data sets affords to the analyst the so called "second order advantage" and the uniqueness of the solution. This is in contrast to the well-known bilinear data where one needs more than one sample to construct the calibration model and the abstract solutions are not unique due to a rotational ambiguity. Various algorithms, such as Parallel Factor Analysis (PARAFAC), Direct Trilinear Decomposition (DTLD) and Positive Matrix Factorization (PMF3), have been created to exploit this feature of the data. However, all of these methods make very na ve assumptions about the characteristics of the noise affecting the data. In order to address this issue, a wide variety of ad hoc variations of these methods (mainly PARAFAC implementations) have been carried out, ranging from pre-scaling to weighted regression. These types of approaches can optimally handle heteroscedastic noise but are inadequate to account for the presence of correlated noise.

Analytical techniques used to yield the multilinear data are prone to generation of correlated noise. Examples include temporal correlation of pump noise in chromatography, spatial correlations of array detectors in spectroscopy, and electronic or digital filters used in many types of instrumentation. In addition to this pernicious and ubiquitous problem, the unfolding/matricization process needed to solve the nonlinear optimization (usually alternating least squares) introduces an additional correlation in another dimension, making the estimation more sub-optimal. The present work introduces and tests a maximum likelihood variant of the PARAFAC algorithm using some simulated data to validate the maximum likelihood properties of the algorithm under a variety of conditions. Also, results obtained from experimental data and comparisons with other methods will be presented. The present algorithm is a natural extension to PARAFAC of the MLPCA method introduced by Wentzell *et al*. It can accommodate heteroscedastic and correlated noise in two or more dimensions and has excellent convergence characteristics because its core is based on an alternating least squares procedure.

Title: The influence of data fusion on multi-way analysis of LC-DAD-MS data

Authors: Ernst Bezemer, Sarah C. Rutan
Department of Chemistry; Virginia Commonwealth University; 1001 West Main Street; Richmond, VA 23284-2006; USA

Keywords: multi-way, LC, DAD, MS, data fusion

Presenter: Ernst Bezemer, drsnooker@yahoo.com

The increased use of multi-hyphenated instruments such as LC-DAD-MS or GC-MS-MS resulting in multi-dimensional data-sets makes multi-way analysis more and more critical for effective utilization of the collected data. Furthermore, when identical samples are analyzed using different forms of instrumentation (such as IR and Raman) or samples are measured in the same instrument over time (while investigating reaction kinetics), these experiments may result in non-linearities due to instrument drift or sample fluctuations.

In theory fusing data from multiple measurements would improve the precision and resolving power of the chemometric methods [1]. However, the very different characteristics of the various data dimensions (for example UV-vis spectra and mass spectra) complicate the analysis. In this study, we investigated the influence of fusing data from various instrumental techniques with our main focus on the fusion of DAD and MS data. The high signal-to-noise data from DAD and the high selectivity from MS are used to aid in the resolution of overlapped liquid chromatography retention profiles [2]. Based on these results, a recommendation for the optimal combination of DAD and MS data is made.

1) P. Wentzell, M. Leger, S. Schreyer, M. Kemper. Maximum Likelihood Principal Components Analysis and Data Fusion: Theoretical and Experimental Investigations Using Near-Infrared and Raman Spectroscopy

2) E. Bezemer, S. Rutan. Study of the hydrolysis of Ally using liquid chromatography with diode array detection and mass spectrometry by three-way multivariate curve resolution. Anal Chem 2001:4403-4409.

Tuesday, September 24

Session 4 *Robust and Graphical Methods*
David L. Duewer, Chair

14:00-14:40 Data Visualization via Sufficient Dimension Reduction
Dennis Cook

14:40-15:05 The Use of Parallel Coordinate Graphical Plotting Combined with Principal Component Analysis Sample Scores for Visualizing Your Data at a Single Glance
Anthony D. Walmsley

15:05-15:30 A Comparative Study of Robust Estimation Methods
Anita Singh

15:30-16:00 Coffee Break

16:00-4:25 Exploring Data Set Structure with Density-based Approaches
Michal Daszykowski

16:25-16:50 Independent Component Analysis and Regression: Applications in Analytical Chemistry
Frank Westad

16:50-17:15 Exploratory Data Analysis of Spectra-structure Similarities
Kurt Varmuza

19:00-22:00 CAC Night at Seattle Mariners

Title: Data visualization via sufficient dimension reduction

Author: Dennis Cook
School of Statistics; University of Minnesota; 1994 Buford Avenue;
St. Paul, MN 55108; USA

Keywords: regression, classification, discrimination, EDA

Presenter: Dennis Cook, dennis@stat.umn.edu

In simple regression a 2D plot of the response versus the predictor displays all the sample information, and can be quite helpful for gaining insights about the data and, if appropriate, for guiding the choice of a first model. Analogous displays of all the data are not possible with many predictors, but fully informative displays are possible in situations where we can find a low-dimensional sufficient summary view that contains or is inferred to contain all the relevant sample information. In this context, regression is defined quite broadly as the study of the conditional distribution of the response given the predictors without pre-specifying a parametric model. Sufficient summary views can be invaluable for guiding the choice of a first model and for gaining insights about the regression, as 2D plots are in simple regression. Seemingly complicated regressions can often be summarized adequately in a relatively simple summary plot.

Foundations of sufficient summary views and methods for constructing them will be discussed. The emphasis will be on regression, but the ideas and some of the methods are equally applicable in other areas like classification and discrimination. Background information is available at www.stat.umn.edu/RegGraph. All of the methods to be discussed are available in the computer program Arc that can be obtained at www.stat.umn.edu/arc. Illustrations will be included in the presentation.

Title: The use of parallel coordinate graphical plotting combined with principal component analysis sample scores for visualizing your data at a single glance

Author: Anthony D. Walmsley
Department of Chemistry; University of Hull; Cottingham Road;
Kingston-upon-Hull HU6 7RX; United Kingdom

Keywords: data visualiazation, process analysis, data reduction, PCA

Presenter: Anthony D. Walmsley, a.d.walmsley@hull.ac.uk

One of the most important tools in chemical analysis is providing visualization that is intuitive, fast, requires little prior data processing and doesn't require extensive training in order to understand the key trends in the data. One method recently gaining a reputation in this area is that of Parallel Coordinates, in which the data from a batch, for example, is displayed on a single plot to allow the users to view the entire data, to check and identify the process bounds, and to ensure the process is functioning as expected. This is a very visual technique, but is limited in the number of variables that can be displayed, and whilst that is not such a problem with typical process data, it limits the use of the method, especially for the chemist.

Typically chemometricians have large quantities of spectral data, and it is simply not feasible to use PC plots to gain any insight with this type of data. However, Principal Components Analysis (PCA) is the classical tool of the chemometrician for data reduction, and so it seem obvious that to combine the PCA approach with parallel coordinates would allow for the analysis such as spectra from a batch or continuous process on a single plot. The beauty of this approach is the simplicity of the plots generated and how easy to understand and spot the trends in the data. A side product of this method is that it not only allows one to view data simply, it also allows the inspection of the PCA process itself, and this makes it much easier to explain the working of the PCA process to the user. It is also possible to add statistical confidence bands to the data, which are an aide to correct interpretation. It is especially useful for identifying outliers in data, as well as allowing easy identification of clusters.

This paper demonstrates the use of parallel coordinates plots with PCA, with examples from laboratory scale, process pilot scale and full process data. The examples show the differences one should typically expect from experimentally designed data, and natural or unplanned process data.

Title: A comparative study of robust estimation methods

Authors: Anita Singh[1], John Nocerino[2]

1 Lockheed Martin Environmental Services; 1050 E. Flamingo Road, Suite E120; Las Vegas, NV 89119-7431; USA

2 National Exposure Research Laboratory; United States Environmental Protection Agency; PO Box 93478; Las Vegas, NV 89193-3478; USA

Keywords: robust estimation, Mahalanobis distance, influence function, confidence interval, Monte Carlo simulation, bias

Presenter: Anita Singh, asingh@lmepo.com

Classical statistical methods often give distorted estimates of the parameters of the main population of interest in the presence of outliers and often fail to identify many of those outliers. In such cases, robust statistical methods are usually chosen to improve the accuracy of the parameter estimates and the outlier identifications. We report on the evaluation of several robust methods to estimate the population mean, , when outliers may be present. Those robust methods are based on: the Huber, Hampel, Biweight, and the proposed (PROP) influence functions; and the trimming method. Some modifications to the Huber and Hampel influence functions defined in terms of the Mahalanobis distance metric are also considered.

The performances of those influence functions are compared using Monte Carlo simulation experiments. Several alternatives to a normal population are assessed to test the ruggedness of the various estimation procedures. The performances of those robust estimation procedures are demonstrated in terms of bias, accuracy ratios, and the 95% nominal coverage probabilities achieved by the respective 95% confidence intervals (CI) of the population mean, . It is observed that the PROP and the trimming methods result in estimates with the least amount of bias and give an average CI length that is close to the expected confidence interval length (ECIL) of the classical method without the outliers. Furthermore, it is shown that the coverage probability of the population mean, , provided by the 95% CI based upon these two robust estimation methods, is close to the nominal 95% coverage.

Title: Exploring data set structure with density-based approaches

Authors: Michal Daszykowski, B. Walczak, D.L. Massart
Farmaceutische en Biomedische Analyse; Vrije Universiteit Brussel; Laarbeeklaan 103; B1090 Brussels; Belgium

Keywords: clustering, density-based methods, data mining

Presenter: Michal Daszykowski, mdaszyk@vub.ac.be

Nowadays collecting and storing analytical measurements are very easy. However this often leads to large and multidimensional data sets, where the information about the investigated chemical phenomenon is hidden. How to extract this information is a very challenging goal for chemometrics, offering a large variety of methods for data structure analysis. Often, the data structure is not homogenous. It means that objects form groups (clusters or patterns) in high dimensional space. Objects located close to each other in the data space, *i.e.* within one group, reflect similar physico-chemical properties.

Thus discovering clusters in the data is considered as the first step in knowledge discovery. For this purpose many methods have been proposed. However, some of them, due to the supervised character and/or required assumptions about data distribution, do not perform well for all kinds of data sets. Density-based methods, that are able to deal with arbitrary shapes of clusters, do.

We focused our attention on two recent density-based algorithms, which fulfill the mentioned properties. Density Based Spatial Clustering of Applications with Noise (DBSCAN) [1,2] and Ordering Points to Identify the Clustering Structure (OPTICS) [3,4] are able to deal with any shapes of data distribution and are only dependent on one input parameter. Being single scan techniques, they process the data during one single pass through the data set, ensuring the computational efficiency. The powerful properties of OPTICS are the possibility to visualize the structure of the multidimensional data on the plane and to judge the discrimination power of the original variables.

1) M. Ester, H. Kriegel, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise, proc. 2nd Int Conf of Knowledge Discovery and Data Mining, Portland, Or, 1996, p. 226, available from www.dbs.informatik.uni-muenchen.de/cgi-bin/papers?query=--CO

2) M. Daszykowski, B. Walczak, D.L. Massart. Looking for natural patterns in data Part 1. Density based approach, Chemom Intell Lab Sys 2001;56:83-92

3) M. Ankrest, M. Breunig, H. Kriegel, J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure available from www.dbs.informatik.uni-muenchen.de/cgi-bin/papers?query=--CO

4) M. Daszykowski, B. Walczak, D.L. Massart. Looking for Natural Patterns in Analytical Data. 2. Tracing Local Density with OPTICS, J Chem Inf Comput.Sci 2002;42:500-507

Title: Independent component analysis and regression: Applications in analytical chemistry

Authors: Frank Westad[1], Lars H. Gidskehaug[2], Harald Martens[3]

1 MATFORSK - Norwegian Food Research Institute; Osloveien 1; N1430 s; Norway

2 Institute of Physical Chemistry; Norwegian University of Science and Technology; N7034 Trondheim; Norway

3 Department of Dairy and Food Science; Royal Veterinary and Agricultural University; Rolighedsvej 30, iii; DK1958 Frederiksberg C; Denmark

Keywords: independent component analysis, source separation, cross validation, regression

Presenter: Frank Westad, frank.westad@matforsk.no

One of the common objectives in data analysis is to extract "pure sources" from a complex set of signals. Principal Component Analysis (PCA) is often applied in chemometrics for exploratory data analysis, and is sometimes combined with a rotation of the axes to interpret underlying structures. However, the extraction of pure and statistically independent spectra from a set of mixtures can not be expected by the use of PCA.

Independent Component Analysis (ICA) has the past years drawn attention due to the potential to extract components that are independent — so-called "blind source separation". Such applications include NIR spectroscopy and compression of images. ICA attempts to recover the original signals by finding a linear transformation that provides statistical independence between the sources under the assumption that the data does not follow a Gaussian distribution. By the use of higher-order statistical information from the densities of the data, this goal may be achieved.

Analytical chemistry is in many situations focused on predicting concentrations of the "pure" compounds and ICA may also be applied in a regression context like Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). These variants of more well-known methods are presented. The importance of cross-validation and significance testing is discussed both for selecting the number of relevant components as well as significant variables. The methods are compared for applications in spectroscopy and electrophoresis.

Title: Exploratory data analysis of spectra-structure similarities

Author: Kurt Varmuza
Laboratory for Chemometrics; Institute of Chemical Engineering;
Vienna University of Technology; Getreidemarkt 9/166; A1060 Vienna; Austria

Keywords: substructure descriptors, similarity of chemical structures, spectral library search, PLS, IR, MS

Presenter: Kurt Varmuza, kvarmuza@email.tuwien.ac.at

Identification of substances or structure elucidation of so far unknown chemical compounds are strongly based on the interpretation of molecular spectra. The most used methods for mass spectra (MS) and infrared spectra (IR) are spectral similarity searches. Result of such library search is a hit list containing the substances of the used spectral library that exhibit spectra that are most similar to that of the unknown. This approach is widely successful in cases the unknown is contained in the library. However, the performance becomes critical when spectra are evaluated from unknowns that are not present in the used library. Some spectroscopic database systems claim an interpretative power if the structures of some hit list compounds are similar to the structure of the unknown. The investigation of structural similarities in hit lists was one aim of this work.

The similarity of spectra has been measured by frequently used criteria mainly based on the correlation coefficient. Chemical structures have been characterized by a set of 1365 binary substructure descriptors calculated by software SubMat. The Tanimoto index was applied to measure the structural similarity. The MS library contained about 100,000 compounds (NIST Mass Spectral Database), the IR library contained about 13,000 compounds (SpecInfo, Chemical Concepts). Random samples with 200 query compounds were used to investigate the degree of similarity between the structures of the hit list compounds and the structure of the unknown. The results were used to compare different approaches for calculating the spectral similarity.

Another aim of this work was to apply PLS for obtaining a better insight into the relationships between spectral data (matrix X with absorbance units for IR or spectral features for MS) and structural data (matrix Y with binary substructure descriptors). PLS-mapping is superior to PCA-mapping and often allows the interactive selection of a compound class (among the hit list compounds) that is relevant for the unknown. A nearest-neighbor search using the PLS-x-scores select relevant compounds in the hit list and thereby facilitates the evaluation of hit lists.

Acknowledgments to W. Demuth, M. Karlovits, H. Scsibrany, and W. Werther for collaboration; S. Stein and R. Neudert for providing the databases. Project P14792 of the Austrian Science Fund.

Wednesday, September 25

Session 5 *Environmental Applications*
Cliff Spiegelman, Chair

8:30-9:10 Utilizing Hourly Gaseous Measurement as an Independent Variable in Multilinear Receptor Model Studies
Philip K. Hopke

9:10-9:35 A N-Way Analysis Technique of Two Tensors Applied to Ozone Concentration Analysis in the Paris Area
Georges Oppenheim

9:35-10:00 Comparison of Factor Analysis Methods for Evaluating the Trace Element Chemistry of Groundwaters of Southern Nevada
Irene M. Farnham

10:00-10:25 Subsampling Particulate Samples: Theoretical Approximations for Environmental Matrices
Robert W. Gerlach

10:25-10:55 Coffee Break

Title: Utilizing hourly gaseous measurement as an independent variable in multilinear receptor model studies

Authors: Philip K. Hopke, Eugene Kim, Pentti Paatero, John Jansen, Eric Edgerton
Department of Chemistry; Clarkson University; Potsdam, NY 13699-5705; USA

Keywords: multilinear engine, receptor modeling, source apportionment, expanded factor models

Presenter: Philip K. Hopke, hopkepk@clarkson.edu

Previously, time resolved wind variables were successfully utilized in multilinear model to reduce rotational ambiguity in receptor modeling [1]. In this study, time resolved gaseous measurement as well as wind data were incorporated into Mutilinear Engine (ME2). We analyzed 24-hour integrated fine particulate matter compositional data measured at Atlanta, GA between August 1998 and August 2000 through the application of this model. A total of 662 samples, 25 compositional variables (20 elements, 3 multi-element ions, and 2 carbonaceous fractions), two wind variables, two calendar variables, and CO measurements were used for this analysis. Preliminary results indicate that time resolved gaseous measurements can be utilized to enhance solutions of multilinear receptor modeling. For most of the local sources, well-defined directional profiles, hourly patterns, seasonal trends, and weekend effects were obtained.

1) P. Paatero, P.K. Hopke. Chemometrics and Intelligent Laboratory Systems 2002;60:25-41.

Title: A N-Way analysis technique of two tensors applied to ozone concentration analysis in the Paris area

Authors: Durville Christophe[1], Franc Alain[2], Georges Oppenheim[1], Saby Claude-Alain[3]

1 Universite d'Orsay; 45 rue Croulebarbe; F75013 Paris; France

2 ENGREF

3 TOTALFINAELF

Keywords: IV Tucker, PARAFAC, cube, ozone pollution

Presenter: Georges Oppenheim, georges.oppenheim@wanadoo.fr

Relatively new methods like Parallel Factor Analysis (PARAFAC), Tucker-models and Multilinear-PLS have been developed [1]. These techniques are designed to decompose higher order data tables, such as cubes of data, in order to reveal the underlying, latent structure for the purpose of data analysis and predictions. These different methods have in common a highly graphical aspect. Drawings are used to determine the position of one sample compared to all others or to evaluate the relative importance of a variable in a multivariate data table. These methods can handle multiple blocks of data collected on the same set of samples, so-called multi-block models. Their use can be beneficial when analyzing large data sets where measurements are organized in conceptually meaningful blocks. The classical approach would be to unfold everything in one big data table and to analyze the entire block. The drawback is a the confusion in the interpretation. Unfolding can be unfavorable for reasons of complexity, interpretation, lost information and poor predictive power.

The one block analysis try to model and to explain the relationship between the different modes. The aim is to determine the fundamental dimensions of the relationship between the modes. The underlying idea is that few dimensions are enough to account for the links. The main information is involved in new modes. These modes are linear arrangements of the levels of the initial modes. For example, the principal components are the linear combinations of the measured variables. In fact each the space generated by the basic variables contains the information.

The multi-block models try to explain the relation between the blocks. The large majority of the scientific papers treat a single set of measurements structured as a cubic form. Nevertheless several authors analyze the case of two data sets [7]. A typical case concerns a tensor Y encoding the chemicals and a tensor B describing the process operating conditions. [Kourti, Nomikos, McGregor]. One of the important industrial framework concerns the batch process where the goals are the monitoring of the trajectories according to satisfactory profiles and the discrimination between normal and abnormal batches.

The simultaneous analysis of two cubes is be dealt with recent works [2,7], and these works don't unfold the tensors: this paper is in keeping of this thought process. The Instrumental Variables approach is used as a complement to the regression variables in order to test the influence of the significant factors. The goal of this paper is to present a technique intended to model the information included in a cube constrained by the information included in a cube. An illustration of the method will be explained on Air Pollution data with a comparison of the results with those of similar methods.

1) C.A. Andersson, R.Bro. The N-way Toolbox for Matlab; Chemom Intell Lab Syst 2000;52:1-4
2) R. Bro. Multi-way analysis in the food industry. Models, algorithms and applications, PhD thesis, University of Amsterdam, 1998.
3) R. Bro. N-PLS, Multiway Calibration. Multilinear PLS, J Chemometrics 1996;10:47-61.
4) A. Franc. Etude Alg brique des multitableaux: apports de l analyse tensorielle. PhD Thesis. Montpellier 1992. 362p.
5) L. Lebart, A. Morineau, M. Piron. Statistique exploratoire multidimensionnelle. Dunod 1995. 439p.
6) S. de Jong, H.A.L. Kiers. Principal covariates regression, Chemom Intell Lab Syst 1992;14:155-164.

7) A.K. Smilde, H.A.L. Kiers. Multiway covariates regression models, J Chemometrics 1999;13:31-48.

Title: Comparison of factor analysis methods for evaluating the trace element chemistry of groundwaters of southern Nevada

Authors: Irene M. Farnham, Ashok K. Singh, Klaus J. Stetzenbach, Kevin H. Johannesson
University of Nevada, Las Vegas; 4505 Maryland Parkway;
Las Vegas, NV 89154-4009; USA

Keywords: PCA, Q-mode FA, correspondence analysis, hierarchical cluster analysis, groundwater, trace elements

Presenter: Irene Farnham, Farnham@UNLV.edu

Numerous investigations have been conducted to characterize the groundwater flow system in the southern Nevada region surrounding the Nevada Test Site (NTS) and the potential repository site for high level nuclear waste at Yucca Mountain. A high degree of uncertainty still exists in the understanding of this groundwater flow system, largely due to the complexity of the geology and the sparse spatial coverage of data from this region. Additional information can be added to the current understanding of the groundwater flow system by evaluating the trace element chemistry of the groundwaters from existing wells and springs of this region. Samples are collected using ultra-clean techniques and analyzed for over 50 different trace elements using an inductively coupled plasma mass spectrometer (ICP-MS). Multivariate statistical techniques are then used to evaluate the large data sets generated.

The multivariate statistical techniques, principal component analysis (PCA), Q-mode factor analysis (QFA), correspondence analysis (CA), and hierarchical cluster analysis (HCA) were applied to a new data set containing trace element concentrations for groundwater samples collected from a number of wells down-gradient from the NTS and also Yucca Mountain, Nevada. The results of each technique reflect the differences in the transformed data matrix used for each of the analyses. PCA results reflect the similarities in the trace element chemistry in the samples resulting from different geochemical processes. QFA results reflect similarities in the trace element compositions and CA is used to evaluate groundwaters based on similarities in the trace elements that are dominant in the waters relative to all other groundwater samples included in the data set.

Title: Subsampling particulate samples: Theoretical approximations for environmental matrices

Authors: Robert W. Gerlach[1], John M. Nocerino[2]

1 Lockheed Martin Environmental Services; 1050 E. Flamingo Road, Suite E120; Las Vegas, NV 89119-7431; USA

2 National Exposure Research Laboratory; United States Environmental Protection Agency; PO Box 93478; Las Vegas, NV 89193-3478; USA

Keywords: particulate sampling, subsampling, representative sampling

Presenter: Robert Gerlach, rgerlach@lmepo.com

Subsampling of environmental samples requires the consideration of many factors. Past sampling practice has relied on classical statistical treatments and, more recently, detailed consideration of the heterogeneous nature of particle matrices. Sampling protocols based on Pierre Gy's sampling theory are discussed and compared to methods developed from less complex, statistical approaches. Environmental particulate samples tend to have different characteristics than those assumed in most published applications of Pierre Gy theory. Common statistical approaches make an entirely different (and sometimes opposite) set of approximations. Critical assumptions for each approach are presented to guide the sampler toward subsampling methods that meet the study objectives and away from practices that result in non-representative results.

Wednesday, September 25

Session 6 *Bioanalytical*
Peter de B. Harrington, Chair

10:55-11:35 Chemometric Opportunities in Proteomics
Alfred L. Yergey

11:35-12:00 Validation of Consensus Between Proteomic Expression and Clinical Chemical Data by a New Randomization F-test in Generalized Procrustes Analysis
Wen Wu

12:00-12:25 Use of Kinetic Equations in Analytical Clinical Chemistry
Jeffrey E. Vaks

12:25-12:50 Modeling the Dynamic Effect of Tea in the Human Body Using Metabonomics - An Exploratory Study
Lefteris Kaskavelis

12:50-14:00 Lunch

Title: Chemometric opportunities in proteomics

Author: Alfred L. Yergey
NIH; Building 10, Room 9D52; Bethesda, MD 20892; USA

Keywords: proteomics, MS

Presenter: Alfred Yergey, aly@helix.nih.gov

The major scientific and technical enterprise widely known as proteomics involves characterizing the protein complement of a cell or organism as a function of physiological state and changes to that state. There is a wide range of methodologies that can be used to address the problem of protein characterization, but to be useful in proteomics any approach must result in an identification of those proteins that have changed in either quantity or extent of modification in response to changes of physiological state. Mass spectrometry is one of the principal tools used in these characterizations. It is used most frequently to identify proteins either by matching a mass spectral fingerprint of peptides produced by enzymatic digestion with *in silico* digests of data base proteins (PMF) or by matching the fragmentation mass spectra of individual peptides using tandem mass spectrometry (MS/MS).

In general, mass spectrometry approaches have been very effective in identifying individual proteins, but somewhat less effective when addressing mixtures. Furthermore, the searching procedures generally used require the protein to be identified be present in a data base. Thus post-translationally modified proteins or those with amino acid substitutions might not be identified since some peptides have been altered. This loss of data is consistent with typical identifications being based on only a fraction, typically 15-20%, of the amino acids in a protein, and also observing that a substantial portion of MS/MS spectra are unused in identifications. Finally, obtaining information from peptides arising from proteins not presently in a database is a class of problems not generally addressed by current methods.

The challenge to chemometrics in proteomics might be stated as taking a fresh look at the field of computer assisted protein identification. One starting point for such an enterprise might be to develop tools for assessing the quality of spectra, particularly MS/MS spectra, prior to attempts to match spectra with predicted fragmentation patterns. In the face of the rapidly growing interest in high throughput mass spectrometric data acquisition, such pre-filtering might yield more rapid analysis times and, more importantly, yield results in which one could place greater confidence than is possible at present. Another potential area for investigation could be the use of pre-filtered MS/MS spectra with standard algorithms to reduce the complexity of data sets by culling peptides from known proteins and then use de novo sequencing approaches to characterize the remaining spectra for true unknowns, modifications or substitutions. The latter approach is illustrated from ongoing work in this laboratory.

Title: Validation of consensus between proteomic expression and clinical chemical data by a new randomization F-test in generalized Procrustes analysis

Authors: Wen Wu, H.C. Cordingley, S.L.L. Roberts, J.R. Armitage, P. Tooke, S.E. Wildsmith
GlaxoSmithKline; The Frythe; Welwyn, Hertfordshire AL6 9AR; United Kingdom

Keywords: proteomics, generalised Procrustes analysis, consensus, validation, randomisation test, significant factors, F-test

Presenter: Wen Wu, Wen_2_Wu@gsk.com

Using proteomic expression data for compound characterization and toxicity prediction has been gathering more and more interest in the pharmaceutical industry over the past few years. However there is no statistical method to assess if the proteomic expression data describes the same toxicological information as the traditional clinical chemical data. In this paper, a new strategy is developed to obtain and validate the consensus between them.

In this strategy, Generalized Procrustes Analysis (GPA) is applied to obtain a consensus between proteomic data and clinical chemical data. The significance of consensus and the dimension of the consensus space are diagnosed by a newly developed method of randomization F-test in GPA (Wu, Guo, de Jong and Massart, Food Quality and Preference 2002, in press). The proposed strategy is applied to match proteomic expression data obtained by SELDI-TOF mass spectrometry to the clinical chemical data in a study of cholestasis in rats.

Two kinds of matching were designed by using animals and treatments as samples in GPA. The results show that the proteomic expression data has significant consensus with clinical chemical data, and that the consensus can be visualized in the group average space with significant factors.

Title: Use of kinetic equations in analytical clinical chemistry

Author: Jeffrey E. Vaks
Beckman Coulter, Inc.; 200 South Kreamer Boulevard, M/S W529;
Brea, CA 92822-8000; USA

Keywords: reaction kinetics, spectroscopic detection, calibration, chemical interferences, error flagging

Presenter: Jeffrey E. Vaks, jevaks@beckman.com

Traditionally, in analytical clinical chemistry (e.g., with spectroscopic detection) reaction kinetics are modeled most often with empirical linear models, sometimes with quadratic polynomial, spline, or polynomial-exponential models. Usually, some rate estimate (initial, peak, in the middle of the window, etc.), obtained with the fitted model, is used for calculating analyte concentration in human fluids. Often, with general clinical chemistry and enzymatic reactions, the relationship between the analyte concentration and rate is practically linear, and calibration is done with single calibrator. With immunoassays involving aggregation of latex particles, the relationship is often described with an S-shaped curve, and calibration requires several levels of calibrator.

The advantage of the empirical modeling of the reaction kinetics is its simplicity. The shortcomings of the empirical modeling are:

- The model approximates the reaction kinetics only in rather short reaction window preventing use of all reaction data available and increasing imprecision.
- The meaning of the model parameters is not clear.
- There are analyte-concentration-dependent biases over the analytical range.
- Error flagging is difficult and involves high rates of the false rejection of reactions providing for acceptable results, as well as of the false acceptance of reactions producing unacceptable results.
- Nonlinear relationship between the analyte concentration and rate requires use of several calibrators increasing cost per test.

Use of kinetic equations for modeling the reactions is more difficult, but it has the potential of overcoming all the above shortcomings of the empirical modeling. The better a kinetic model describes the reaction, the wider is the reaction window it fits to and the higher is the precision, and the smaller are the biases. Knowledge of the meaning of the model parameters allows for using certain functions of the model parameters that are in linear relationship with the analyte concentration, which allows for single-point calibration. It also allows for meaningful flagging the reactions that deviate from the established kinetics. Also, sensitivity of the analytical results to certain types of chemical interferences, that are included in the kinetic model, can be minimized. Some examples of kinetic models and their potential in analytical clinical chemistry are discussed.

Title: Modeling the dynamic effect of tea in the human body using metabonomics - An exploratory study

Authors: Lefteris Kaskavelis, Clare A. Daykin, Hai Pham Tuan
Unilever Research & Development Vlaardingen; PO Box 114;
NL3130 AC Vlaardingen; The Netherlands

Keywords: metabonomics, smoothing splines, wavelets, HPLC-DAD, LC-MS, PARAFAC, batch analysis

Presenter: Lefteris Kaskavelis, Lefteris.Kaskavelis@unilever.com

The area of metabonomics has recently been introduced as a branch of biochemistry where multivariate statistical methods and analytical measurement techniques are coupled for modelling the time-dependent effects of metabolites within an organism. In this paper a number of statistical methods for data compression, pattern recognition and feature extraction are used for achieving two aims. First, modelling the dynamic effect of the tea dose to humans and subsequently identifying key areas in the analytical signals that are responsible for changes in the metabolic profile.

Human urine samples were analysed by means of HPLC-DAD and LC-MS resulting in data sets with a 3D structure. A combination of wavelets and the PARAFAC model for data compression and modelling is shown to give good results in terms of information extraction.

Subsequently smoothing techniques based on B-splines are applied on the derived scores to model the dynamic characteristics of the metabolic response due to the tea dose. By modelling the data as a sequence of batches, starting at the time of the tea dose, the magnitude and the duration of the tea effect could be estimated. The loadings plots are used to identify areas (peaks) in the analytical signals responsible for the metabolic changes due to the tea. Finally application of LC-MS-MS has given some preliminary information on identifying compounds of interest.

Wednesday, September 25

Session 7 Genetic Algorithms, Neural Networks and Datamining
Steven D. Brown, Chair

14:00-14:40 Support Vector Machines for the Classification of Electronic Nose Data
Matteo Pardo

14:40-15:05 Selection of the Optimal Inputs in Chemometrics Modeling by Artificial Neural Network Analysis
Zvi Boger

15:05-15:30 Real-time Chemometrics Applied for Screening Food-Borne Pathogens and Bacterial Biomarker Using Ion Mobility and Differential Mobility Spectrometries: Chemometrics^{N}
Peter de B. Harrington

15:30-16:00 Coffee Break

16:00-16:25 Growing Neural Networks for Feature Selection and Calibration of Sensor Set-ups
Frank Dieterle

16:25-16:50 The Use of Continuous and Discrete Variables for Regression and Classification in Bayesian Networks
Nathaniel A. Woody

16:50-17:15 The O-PLS Approach, a New Modeling Concept in Multivariate Calibration
Johan Trygg

18:00-22:00 Conference Dinner at Odyssey, the Maritime Discovery Center

Title: Support vector machines for the classification of electronic nose data

Authors: Matteo Pardo, G. Sberveglieri
INFM & University of Brescia; Via Valotti, 9; I25133 Brescia; Italy

Keywords: support vector machines, classification, kernel, multilayer perceptrons, electronic nose

Presenter: Matteo Pardo, pardo@tflab.ing.unibs.it

In this contribution we introduce Support Vector Machines (SVM) and apply them to the classification of two binary datasets of different hardness. The datasets consist of electronic nose (EN) measurements of different types of coffee blends.

SVM emerged in the late 90s from statistical learning theory (SLT) as an efficient classification algorithm and has been successfully applied to a number of problems ranging from face identification and text categorization to bioinformatics and data mining. In short, the idea of SVM is:

- Map input vectors non-linearly into a high dimensional feature space.
- Construct the optimal separating hyperplane in the feature space. The optimal hyperplane is the one that maximizes the margin.

The advantages of SVM are:

- It is theoretically well founded, being based on SLT. Large margin hyperplanes have small VC dimension and hence the generalization error is bound more tightly.
- It is practical, as it reduces training to a quadratic programming problem with a unique solution. By introducing a kernel function, all computations are directly performed in the input space. The dimensionality of the feature space is not a computational issue.
- The resulting classifier depends only on a subset of the training data, the so called support vectors. Therefore SVM can be used for data compression.
- It contains a number of heuristic algorithms as special cases: by the choice of different kernel functions, we obtain e.g. polynomial classifiers, RBF classifiers and MLP.

We investigate the performance of SVM on the EN datasets with regard to:

- The number of principal components (PC), that are given as inputs to the SVM. These range from two to five, five being the number of sensors used in the EN.
- The type of kernel: we tried both polynomial and RBF kernels. For the polynomial kernel, we examined polynomial of various order while for RBF kernels we scanned through different variance parameters. The performance for each parameter was measured by 10 fold cross-validation.

We noticed that both kernels led to similar performance and to strongly overlapping sets of support vectors. Also a comparison to traditional multilayer perceptrons (MLP) has been undertaken, where early stopping and the Levenberg-Marquardt algorithm have been considered for training MLP. SVM outperformed MLP though the difference was not significant. Further, the dependence of the classification error on the number of PC wasn't critical for SVM, while the contrary is the case for MLP classification.

Title: Selection of the optimal inputs in chemometrics modeling by artificial neural network analysis

Author: Zvi Boger
OPTIMAL - Industrial Neural Systems Ltd.; 261 Congressional Lane, Suite #319; Rockville, MD 20852; USA

Keywords: artificial NNs, instrumentation spectra, input selection

Presenter: Zvi Boger, zboger@bgumail.bgu.ac.il

Instrumentation spectra used for chemometrics analysis are often unwieldy to model, as many of the inputs do not contain important information. Several mathematical methods are used for reducing the number of inputs to the significant ones only.

Artificial neural networks (ANN) modeling suffers from difficulties in training models with a large number of inputs. However, the PCA-CG non-random initial connection weight algorithm can overcome these difficulties [1-2]. Once the ANN model is trained, the analysis of its connection weight can easily identify the more relevant inputs [3]. Repeating the process of training the ANN model with the reduced input set and the selection of the more relevant inputs can proceed until an optimal, small, number of inputs can be identified.

These ANN techniques have been already used for chemometrics modeling [4,5]. Results of recent work will be presented, including the modeling of artificial nose ; sensor array data with 1000+ inputs that were reduced to optimal sets of less than 10 inputs [6]. The accuracy of the resulting ANN models is usually better, and more robust, than the original large ANN model.

1) H. Guterman. Neural, Parallel and Scientific Computing 1994;2:43-54.

2) Z. Boger. Information Sciences - Applications, 1997;101(3:4):203-212.

3) Z. Boger, H. Guterman. Proc, IEEE Int Conf on Systems Man and Cybernetics, Orlando, 1997, 3030-3035.

4) Z. Boger, Z. Karpas. Anal Chim Acta 1994;292:243-251.

5) Z. Boger, Z. Karpas. J Chem Inf Comput Sci 1994;34:576-580.

6) Z. Boger, S. Semancik, R.E. Cavicchi, Proc. 9th Int Meeting on Chemical Sensors, Boston, July 2002.

Title: Real-time chemometrics applied for screening food-borne pathogens and bacterial biomarker using ion mobility and differential mobility spectrometries: Chemometrics^N

Authors: Peter de B. Harrington, Guoxiang Chen, Libo Cao
Department of Chemistry & Biochemistry; Clippinger Labs; Ohio University; Athens, OH 45701-2979; USA

Keywords: bacteria, bioterrorism, ion mobility spectrometry, differential mobility spectrometry, multidimensional compression, wavelet, NN, LDA, SIMPLISMA, chemometrics^N

Presenter: Peter de B. Harrington, Peter.Harrington@Ohio.edu

A rapid and portable method for detecting pathogenic bacteria is desired for maintaining hygiene standards in food processing plants and preventing outbreaks of food poisoning. A recent estimate by the Centers for Disease Control of food borne bacterial infection is 76 million cases per year for the USA that resulted in 5,000 fatalities [1]. Concerns have arisen regarding Homeland Security and safeguarding the food supply against bioterrorist attack. Some food poisoning cases, such as Scombrotoxic poisoning, are caused by the degradation of meat and the production of biogenic amines. Some biogenic amines (e.g. cadaverine and putrescine) may provide easy to identify indicators of food spoilage and pathogenic bacterial activity.

Ion and differential mobility spectrometries (IMS & DMS) afford low cost and portable instruments. A micro-machined DMS instrument is less than 2 cm in length and is amenable to the design of handheld spectrometers. These instruments are ideally suited for amine detection. The charge transfer chemistries inherent to IMS and DMS are intricate, and when the instruments are used outside the controlled environment of the laboratory or used with complex samples such as whole cell bacteria, the instrumental response may become difficult to interpret.

SIMPLISMA [2] has been implemented in a real-time system for simultaneously acquiring and modeling IMS data [3]. The SIMPLISMA spectra represent concentration independent spectra that can be used for classification models constructed using classical approaches such as linear discriminant analysis [4] or modern approaches such as temperature constrained neural networks [5]. Multidimensional wavelet compression [6] has been used to compress IMS data to manageable sizes (*i.e.,* less than 1%). The compressed data is modeled with SIMPLISMA. The SIMPLISMA models are inverse transformed to provide a global perspective of the compressed data [7]. This paper will present an overview of how a combination of these chemometric methods enables the use of IMS [8] and DMS on complex biological samples.

1) http://www.cdc.gov/communication/tips/foodborne.htm (Date Accessed 25 April).
2) W. Windig, J. Guilment. Interactive Self-Modeling Mixture Analysis. Anal Chem 1991;63:1425-1432.
3) G. Chen; P.D. Harrington. Real-time interactive self-modeling mixture analysis. Appl Spectrosc 2001;55:621-629.
4) A. Mehay, C.S. Cai, P.D. Harrington. Regularized Linear Discriminant Analysis of Wavelet Compressed Ion Mobility Spectra. Appl Spectrosc 2002;15:219-227.
5) P.D. Harrington. Temperature-constrained cascade correlation networks. Anal Chem 1998;70:1297-1306.
6) A.A. Urbas, P.D. Harrington. Two-dimensional wavelet compression of ion mobility spectra. Anal Chim Acta 2001;446:393-412.
7) P.D. Harrington, P.J. Rauch, C.S. Cai. Multivariate curve resolution of wavelet and Fourier

compressed spectra. Anal Chem 2001;73:3247-3256.

8) P.D. Harrington, T.L. Buxton, G. Chen. Classification of Bacteria by Thermal Hydrolysis Methylation Ion Mobility Spectrometry and SIMPLISMA. Int J Ion Mobil Spectrom 2001;4:148-153.

Title: Growing neural networks for feature selection and calibration of sensor set-ups

Authors: Frank Dieterle, Birgit Kieser, Stefan Busche, G nter Gauglitz
Institut fuer Physikalische und Theoretische Chemie; Universit t Tuebingen;
Auf der Morgenstelle 8; DE72076 Tuebingen; Germany

Keywords: growing NNs, feature selection, time-resolved measurements

Presenter: Frank Dieterle, frank.dieterle@ipc.uni-tuebingen.de

The classical approach of quantifying several analytes in mixtures is the combination of a number of sensors on an array, whereby the sensors are supposed to show different sensitivities for the analytes to be investigated. In most cases, one feature per sensor is used for the calibration by a multivariate method. This classical approach is limited by the need of as many sensors as analytes to be quantified.

In this work the time-resolved information of the sensor response is exploited to overcome this limitation. A nonlinear sensor response of a single sensor is recorded at a number of time points and used like a huge virtual sensor array . In order to identify the most significant time points and to prevent an over-fitting feature selection is needed. In this study the calibration and feature selection is performed by the use of growing neural networks. The algorithm starts with an empty neural network and tries to minimize the sum square error of prediction by successively inserting units with input and output links or by inserting pure links. The algorithm calculates the reduction of the error by retraining the network after the insertion of the new element to find the best place for the insertion.

The complete procedure uses many parallel runs of this growing network algorithm on different training and test sets generated by a multiple holdout procedure. In a second step, the time-points are ranked according to the frequency of usage in the parallel runs. Then the time-points are added to a neural network until the improvements in the prediction are not significant any more.

The work presented here focuses on a simultaneous quantification of methanol ethanol and propanol in ternary mixtures using the time-points of only one sensor of a SPR set-up. The proposed algorithm outperforms the calibration and prediction of neural networks using all time points.

Title: The use of continuous and discrete variables for regression and classification in Bayesian networks

Authors: Nathaniel A. Woody, Steven D. Brown
Department of Chemistry & Biochemistry; University of Delaware;
Newark, DE 19716; USA

Keywords: Bayesian networks, classification, missing data

Presenter: Nathaniel Woody, warsaw@udel.edu

The use of Bayesian statistics and Bayesian systems has become widespread over the past 10 years in many areas of computer science, social science, and in the rapidly growing informatics fields, including areas like genomics and proteomics. Bayesian classifiers have been demonstrated to perform well in many situations; here we demonstrate how to expand beyond the simple Bayesian classifier into a Bayesian network that allows a fuller representation of a data domain. The probabilistic nature of a Bayesian network also allows an interpretation of a network as a methodology for handling missing data. The means for creating networks that handle missing data are demonstrated as simple extensions of discrete Bayesian classifiers. This extension is accomplished by a feature-selective structure learning mechanism that builds in estimators for missing values. The resulting structures are applied to chromatographic data to demonstrate the ability to create a classification model that is robust to missing data in both training and test cases.

Extending this methodology for handling missing data to continuous variables is difficult within the framework of the standard Bayesian network representation. The standard representation of continuous variables has relied on an MLR-like regression model to propagate means and covariances through a generalized linear model. This method tends to produce poor regression models in situations with weakly correlated variables. To address this problem PCR and PLS are used to produce regression vectors in the Bayesian network. The ability to transfer variance information through a Bayesian network is examined using standard statistical datasets to demonstrate the ability of the network to perform regression and filling operations.

Title: The O-PLS approach: A new modeling concept in multivariate calibration

Authors: Johan Trygg, Svante Wold
Ume University; SE901 87 Ume ; Sweden

Keywords: parsimonious multivariate calibration models, structured noise, O-PLS approach, PLS, model interpretation

Presenter: Johan Trygg, j.trygg@imb.uq.edu.au

Spectroscopic (e.g. NMR, NIR) and chromatographic techniques (e.g. GC, LC) are frequently being used for the characterization of solid, semi-solid, fluid and vapor samples. Multivariate calibration methods (e.g. partial least squares projections to latent structures (PLS)) are often used to develop a quantitative relation between the digitized spectra, the matrix X, and some properties (e.g. concentrations) of the analytes, the matrix Y. These methods may also be used to infer other more multivariate properties of samples, e.g. predicting the NMR profiles from NIR spectra. This large quantity of information-rich (not necessarily all relevant) data requires proper multivariate tools. Examples of non-relevant systematic variations in spectroscopy are baseline and scatter effects, as well as spectra of impurities or unknown constituents. These adversely affect the interpretation of the PLS (and other methods with similar properties) model and increase model complexity. Pre-processing methods (e.g. orthogonal signal correction filters) can be applied to suppress this structured noise. However, these methods do not take the calibration model into account.

Here, we describe the O-PLS approach [1-3]. O-PLS represents a new concept in two-block (X-Y) modeling because instead of mixing all variation together in the prediction model, it explicitly separates a) the Y-orthogonal variation in X, b) the X-orthogonal variation in Y and c) the joint X-Y covariation. The general O-PLS model is,

Model of X: $X = TW' + TyoPyo' + E$
Model of Y: $Y = UC' + UxoPxo' + F$
Prediction of Y: $Yhat = TC'$

where some factors (T) are common to both X and Y.

Examples (synthetic and real) will demonstrate how the PLS model and its parameters (e.g. scores and loadings) are adversely affected when structured noise is present in X and Y. Comparison with O-PLS will demonstrate the versatility and ability of this new approach to produce interpretative, parsimonious prediction models with the same predictive ability as PLS.

1) J. Trygg, S. Wold. Orthogonal projections to latent structures, O-PLS. J Chemometr 2002;16(3):119-128.

2) J. Trygg. O2-PLS for qualitative and quantitative analysis in multivariate calibration. J Chemometr 2002;16(6):283-293.

3) J. Trygg. Parsimonious Multivariate Models. PhD thesis, Ume University, 2001. http://www.chem.umu.se/dep/orgchem/forskning/thesis/johantrygg/jtabstract.stm

Thursday, September 26

Session 8 Calibration
Anthony D. Walmsley Chair

8:30-9:10 Wavelet Multiscale Regression Analysis for Multivariate Calibration
Steven D. Brown

9:10-9:35 Multivariate Calibration With Incomplete Designs
Clifford H. Spiegelman

9:35-10:00 Fast Algorithm for the Solution of Inequality Constrained Least Squares Problems
Mark Van Benthem

10:00-10:25 Transformation of Sensor Array Pattern Vectors into Descriptors of Unknown Vapors using Classical Least Squares and Inverse Least Squares Methods
Jay W. Grate

10:25-10:55 Coffee Break

10:55-11:20 Importance of Spectral Errors on Predictions Obtained by Partial Least Squares Regression
Charlotte M¿ller Andersen

11:20-11:45 Improving Piecewise Orthogonal Signal Correction
Huwei Tan

11:45-12:10 Calibration and Instrumental Design Strategies for Physiological Glucose Measurements Based on Near-IR Spectroscopy
Gary W. Small

12:10-13:10 Lunch

Title: Wavelet multiscale regression analysis for multivariate calibration

Authors: Steven D. Brown, HuWei Tan
Department of Chemistry & Biochemistry; University of Delaware;
Newark, DE 19716; USA

Keywords: wavelets, multivariate calibration, multi-scale

Presenter: Steven D. Brown, sdb@udel.edu

Well-established algorithms based on direct application of regression by partial least squares (PLS) or principal component regression (PCR) are the most widely used methods for multivariate calculation. They explain global spectral variance by using latent variables in a single-scale representation. As a result, more latent variables have to be used to explain local sources of variance, leading to unnecessarily complicated calibration models. The spectral vectors in a calibration can be viewed in a multi-scale way, since spectral signals contain contributions with different frequency bands and at different wavelengths from a variety of sources, such as detector noise, instrumental differences, temperature effects, and sample variations. The spectral vectors can be interpreted as summations of the frequency components over the wavelet frequency domain. Plots of the spectral entropy versus scale shows that the spectral variations related to the properties of interest appear mostly in the middle range of frequencies. The multi-scale representation of spectral signals over the wavelet frequency domain provides us with a way to improve calibration models by isolating the irrelevant variation in a set of spectra.

Our new method, wavelet multiscale regression analysis (WMRA), is intended to simplify the complexity and to improve the performance of multivariate calibration models. It is a two-step procedure, conducted in a way similar to calibration using regular regression methods. The first step is to establish and optimize a multiscale model in a calibration set between the dependent m « 1 vector y (property) and independent multiscale spectra tensor X {X_k, k = 1, 2, , L+1} by means of the multiscale regression model

$$y = \text{sum}(k=1 \text{ to } L+1)\ X_k\ b_k + e$$

where E(e)=0, Cov(e)=sigma2 and where bk is the p x 1 regression coefficient vector for the frequency component at the kth scale in multiscale spectra, e denotes an m « 1 error vector, and E(Æ) and Cov(Æ) are the expectation and covariance, respectively. The second step is to predict values for the dependent properties based on an independent test set.

The goal of the multiscale regression analysis is to calculate the multiscale regression coefficients b = {b_1, , b_{L+1}} with the lowest prediction errors. PCR, PLS, ridge regression and conventional multiple linear regression using the maximum likelihood criterion or the Bayesian information criterion are suitable approaches for the regression step. Our results suggest that significant improvement in prediction errors can often be achieved as compared to conventional PCR/PLS with or without conventional wavelet processing.

Title: Multivariate calibration with incomplete designs

Authors: Clifford H. Spiegelman[1], Sang-Joon Lee[1], Joseph M. Conny[2], Frits H. Ruymgaart[3]

1 Department of Statistics; Blocker Building; Texas A&M University; College Station, TX 77845-3143; USA

2 Surface Science and Microanalysis Division; National Institute of Standards and Technology; 100 Bureau Drive, Stop 8372; Gaithersburg, MD 20899-8372; USA

3 Texas Tech

Keywords: classical and inverse calibration

Presenter: Clifford Spiegelman, chemometrics@aol.com

There has been some debate whether inverse or classical calibration methods are superior when there are multivariate predictors and some of them are missing. In this paper we compare these two methods in the case where the design is not completely known. We develop some general results in the multivariate case and carry out extensive simulations in a univariate model with partly known regressors and several error distributions. These simulations reveal that the methods perform differently, depending on the specifics of the model. Neither method, however, turns out to be consistently superior to the other when the rank of the spectra is smaller than the number of ingredients in the physical sample.

Title: Fast algorithm for the solution of inequality constrained least squares problems

Authors: Mark H. Van Benthem, Michael R. Keenan
Sandia National Laboratories; Mail Stop 0886; Albuquerque, NM 87185-0886; USA

Keywords: least squares, MCR-ALS, constrained least squares

Presenter: Mark Van Benthem, mhvanbe@sandia.gov

Algorithms for multivariate image analysis and other large-scale applications of multivariate curve resolution (MCR), typically employ constrained alternating least squares (ALS) procedures in their solution. The solution to a least squares problem under general linear equality and inequality constraints can be reduced to the solution of a non-negativity constrained least squares (NNLS) problem. The efficiency with which the constrained least squares problems can be solved, therefore, rests heavily on the underlying NNLS algorithm. We will present a new NNLS solution algorithm that is appropriate to large-scale MCR and other ALS applications. This new algorithm rearranges the calculations in the standard active set NNLS method on the basis of combinatorial reasoning. This rearrangement serves to reduce substantially the computational burden required in a large-scale MCR problem.

Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

Title: Transformation of sensor array pattern vectors into descriptors of unknown vapors using classical least squares and inverse least squares methods

Authors: Jay W. Grate[1], Barry M. Wise[2], Neal B. Gallagher[2]

1 Environmental Molecular Sciences Laboratory; Battelle Pacific Northwest National Laboratory; PO Box 999, Mail Stop K8-93; Richland, WA 99353; USA

2 Eigenvector Research, Inc.; PO Box 561; Manson, WA 98831; USA

Keywords: sensor array, pattern, classical least squares, inverse least squares, descriptor

Presenter: Jay W. Grate, jwgrate@pnl.gov

The prevailing approach in the use of sensor arrays for vapor identification is that the array must be trained to recognize the vapor or vapors of interest. In this essentially empirical approach, compounds not in the training set cannot be identified. The sensor array, however, is collecting multivariate chemical information about the sample. In principle, information should be extractable from the array response to gain knowledge of the chemical properties of the unknown compound, whether that particular compound was in the training or not.

Polymer-coated acoustic wave sensors represent a sensor technology that is particularly well characterized in terms of the sensors transduction mechanisms and the interactions of analyte species with the polymeric sensing layers. Examples of acoustic wave devices used in chemical sensing applications include the thickness shear mode (TSM), surface acoustic wave (SAW), and the flexural plate wave (FPW) device. Acoustic wave vapor sensors respond to any vapor that is sorbed at the sensing surface with a response that is proportional to the amount of vapor sorbed.

In this paper we will demonstrate how inverse least squares (ILS) and classical least squares (CLS) approaches can be applied to data from a well-understood polymer-coated acoustic wave vapor sensor array to transform sensor array responses into values of descriptors of the detected vapors. Using SAW array responses and ILS models, five descriptors are determined for each of 18 organic vapors, and these are compared with the known descriptor values. Calibration and prediction results are examined, as well as various ILS regression models.

Title: Importance of spectral errors on predictions obtained by partial least squares regression

Authors: Charlotte M¿ller Andersen, Rasmus Bro
Department of Dairy and Food Science; Royal Veterinary and Agricultural University; Rolighedsvej 30, iii; DK1958 Frederiksberg C; Denmark

Keywords: PLS, spectral errors, replicates

Presenter: Charlotte M¿ller Andersen, cma@kvl.dk

Application of spectroscopy for quality measurements in the food industry is a promising area where non-destructive and cheap measurements can replace more complicated, expensive or dangerous measurements of the product. However, a spectral measurement is often to represent a whole object even though only a small part of it is actually measured. For example measurements performed on a localized area of an object are used as a measurement representative of the whole object. This may introduce an error due to the inhomogeneity of the product and together with other errors resulting from the measuring process this influences the calibrations.

Partial least squares regression (PLS) is used for predicting quality measurements (reference measurements) from the spectral data. Normally, the performance of the predictions are validated by calculating the RMSEP giving only one error estimate for each prediction. In addition to the errors in the spectral measurements, there will be errors in the reference measurements as well as errors originating from the modeling step. All these errors will contribute to the total error of the predictions and the importance of the single contribution will depend on various factors such as the size of the error, number of replicates, etc.

An errors-in-variables approach is used for analyzing the effect of the various types of errors on the predictions. Error propagation is used for deriving expressions that also accounts for errors in the independent variables. The purpose is to investigate when the error in the spectral measurements is important, and in case of importance how to minimize the influence of this error.

The effect of error in the spectral data can be reduced using the average of replicated X-measurements both for modeling and for prediction. However, the number of replicates to use is often chosen as a compromise between the precision required and the cost of making the measurements. Therefore, a second purpose concerns investigating the effect of replicates and finding the optimal number of measurements to use in the various situations of error contributions.

Title: Improving piecewise orthogonal signal correction

Authors: Huwei Tan, Robert N. Feudale, Steven D. Brown
Department of Chemistry & Biochemistry; University of Delaware;
Newark, DE 19716; USA

Keywords: OSC, piecewise OSC, signal processing, PLS

Presenter: Huwei Tan, hwtan@udel.edu

Piecewise orthogonal signal correction (POSC) that performs OSC-like correction in a piecewise manner, was developed recently in this research group to process spectral signals. The work will be presented here carries this POSC algorithm one step further. Based on a better understanding of the linear algebra for the POSC and OSC, the algorithm has been meaningfully simplified, where the spectral correction is performed by an eigendecomposition and a direct orthogonalization in a piecewise manner. The improved POSC is applied to two near-infrared (NIR) data sets for multivariate calibration and the results compared with those obtained by the current versions of the OSC algorithms. It is shown that performing the improved algorithm prior to calibration yields regression models that are more parsimonious (fewer latent variables) and with better predictive power than models obtained with OSC.

Title: Calibration and instrumental design strategies for physiological glucose measurements based on near-IR spectroscopy

Author: Gary W. Small
Department of Chemistry & Biochemistry; Clippinger Labs; Ohio University;
Athens, OH 45701-2979; USA

Keywords: NIR, glucose, calibration, interferometer, optical filter

Presenter: Gary Small, small@ohio.edu

The determination of blood glucose levels by near-infrared spectroscopy offers the potential of noninvasive, continuous monitoring of this clinically important blood constituent. As a replacement for current invasive glucose home testing procedures, this measurement capability would provide significant benefits to diabetic patients, in terms of both better management of their disease and improved quality of life. The technology would also find use in hospital settings for applications such as patient bedside monitoring.

Two issues that impede the successful development of near-infrared blood glucose measurements are the implementation of a successful and stable calibration model to relate the spectral measurements to the glucose level and the need to produce a simple, rugged instrument for the measurement that still possesses the required optical performance. These two issues are intertwined because the calibration requirements dictate the characteristics of the measurement platform in terms of parameters such as spectral range, signal-to-noise ratio, scan speed, and resolution.

In this presentation, both of these issues will be addressed in the context of glucose measurements made in two model systems that are designed to simulate the pertinent characteristics of the physiological measurement. These systems are:

1) glucose in an aqueous buffer consisting of variable levels of bovine serum albumin (simulant for human blood proteins) and triacetin (simulant for triglycerides) and
2) glucose in an aqueous buffer containing variable concentrations of lactate, alanine, urea, triacetin, and ascorbate.

Specific topics to be addressed include calibration strategies for overcoming instrumental drift and ways to use the results of calibration experiments to simplify the instrumental design. Two instrumental designs are simulated in this work:

1) a short-scan interferometric system in which the calibration is based on digitally filtered interferogram segments and
2) a filter-based system in which the optical filters encode the multivariate components required to implement a calibration model.

Poster Program Abstracts

Title: Automatic preprocessing of LC-MS data

Authors: Fredrik O. Andersson, Sven P. Jacobsson
AstraZeneca R&D; Forskargatan 20; SE151 85 Sodertalje; Sweden

Keywords: LC-MS preprocessing

Presenter: Fredrik O. Andersson, fredrik.o.andersson@astrazeneca.com

Datasets produced from LC-MS systems are usually quite big and to extract useful information can be a daunting task, especially when it's from a complex mixture. In this study we present routines (in Matlab) for automatic extraction from the CDF-files (exported from the LC-MS system) and preprocessing of such a data.

Initially, chromatographic profiles for a specific mass number, starting with the mass number with the highest intensity are extracted. This continues until a set threshold is reached. A screening function to only select the valid profile, profiles containing real fragments and not solvent, noise, etc. Any profile in the matrix containing more then one peak is spliced. Profiles that have their maximum intensity in the same given retention time window, are set to belong to the same substance. The mass to charge numbers for a found (given) substance is then extracted from the entire sample series to produce a structure with mass to charge numbers, retention time and intensity for each substance.

The classification and regression properties of preprocessed data is illustrated by a sample subjected to accelerated degradation conditions.

Title: A multivariate screening approach using statistical experimental design and PLS for metabonomic NMR and clinical chemical toxicity data

Authors: Henrik Antti, Tim Ebbels, Hector Keun, Mary Bollard, Olaf Beckonert, Elaine Holmes, John Lindon, Jeremy Nicholson
Biological Chemistry, Biomedical Sciences Division; Imperial College of Science, Technology & Medicine; Sir Alexander Fleming Building, Exhibition Road; South Kensington, London SW7 2AZ; United Kingdom

Keywords: DoE, PLS, NMR, metabonomics, biomarker screening, COMET

Presenter: Henrik Antti, h.antti@ic.ac.uk

Metabonomics is increasingly recognized as a valuable approach for investigating metabolic changes in biofluids and tissues caused by drug toxicity, or genetic intervention. It uses ^{1}H-NMR spectroscopy to derive the data and because extensive data sets are required consideration has to be given to optimum design of experiments as for subsequent data analysis.

The Consortium for Metabonomic Toxicology (COMET) is an academic project supported by the pharmaceutical industry that aims to construct databases of drug toxicity using ca. 100,000 ^{1}H-NMR spectra of biofluids from laboratory rats and mice treated with model compounds, using chemometric methods to characterize the time-related and dose-specific multi-parametric responses of toxins on their endogenous metabolite profiles.

Here we present a multivariate approach to screening for statistically significant markers of toxicity based on statistical experimental design (DoE) and partial least squares regression (PLS). The methodology was applied to COMET-data from two toxicology studies in the rat including the model toxins a-naphthylisothiocyanate (ANIT) and hydrazine.

For the ANIT study, 600 MHz ^{1}H-NMR spectra of urine samples from control rats and dosed rats (125 mg/kg) collected at 8 and 48 h post-dose were processed and data reduced according to standard metabonomic protocols. For the hydrazine study, 2D J-resolved (J-RES) ^{1}H-NMR spectra and the corresponding clinical chemistry measurements of blood serum samples from control rats and dosed rats (30 and 90 mg/kg) collected at 48 and 168 h post-dose were analyzed.

The PLS analysis provided an interpretation of the correlation structure within and between the data blocks (NMR data and clinical chemistry measurements) according to dose, time and the interaction between the two (dose x time). Calculation of confidence intervals for the PLS regression coefficients allowed a classification of variables according to statistical significance for the changes in each spectral area and clinical chemistry parameter related to dose and/or time.

Our suggested DoE-PLS approach provided an efficient means for biomarker screening and detection of toxicity-related patterns in metabonomic NMR data based on statistical significance. The method also facilitated analysis of multiple blocks of multiparametric data suggesting its value for other types of complex biological data such as generated within the fields of proteomics and genomics.

Title: Combined *in vivo* transcriptomics and metabonomics in characterization of biomarkers for mechanistic studies of lipid-lowering drugs

Authors: Dorrit Baunsgaard, Klaus S. Frederiksen, Jan Fleckner, Erik Max Wulff, Karsten Wassermann, Per Sauerberg, Ulla G. Sidelmann
Novo Nordisk A/S; Novo Nordisk Park; DK2760 Maaloev; Denmark

Keywords: transcriptomics, metabonomics, microarray, NMR, biomarkers, variable selection, dose-response model correlation

Presenter: Dorrit Baunsgaard, doba@novonordisk.com

Transcriptomics, proteomics and metabonomics are rapidly developing technologies that enable researchers to study and describe biological events at different levels in organisms. Expression can be studied at the stage of transfer of genetic information (transcriptomics), at the stage of formation of proteins (proteomics), and by determining the metabolites resulting from the activities of those proteins (metabonomics). The data generated by these methods can be used to characterize changes in potentially thousands of biomolecules simultaneously. The large amounts of data, demand the use of multivariate statistical analysis to extract the information of interest, and the usefulness of linear methods such as principal component analysis (PCA) and partial lest squares regression (PLSR) for classification purposes has already been established [1,2].

In the pharmaceutical industry there is a huge interest in the use of these technologies for the identification of biomarkers (genes, proteins and metabolites) in safety and efficacy evaluations of novel disease interventions. To ensure that the biomarkers at the different biomolecular levels are all related to the same biological changes, data reduction of the data sets of the various technologies should yield models that consequently also correlate to each other. In the present study we applied transcriptomics (cDNA microarrays) to tissue samples and metabonomics (NMR spectroscopy) to plasma samples of high fat fed rats treated with lipid-lowering drugs. We used PLSR followed by variable selection on the microarray data and the NMR data, respectively, to find biomarkers that varied in relation to the dose-response conditions of the drugs. The dose-response related gene and metabolite biomarkers were then combined in a new PLSR model to find the direct correlation between individual biomarkers at both gene and metabolite level to elucidate pathways involved in the drug effect.

1) F. Jessen, R. Lametsch, E.R. Bendixen, I.V.H. Kj rsg rd, B.M. J¿rgensen. Extracting information from two-dimensional electrophoresis gels by partial least squares regression. Proteomics 2002;2:32-35.
2) J.K. Nicholson, J. Conelly, J.C. Lindon, E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. Nature Reviews Drug Discovery 2002;1:153-161.

Title: Building NMR spectroscopic metabonomic databases: HCA and KNN classification for the exploration and prediction of drug toxicity

Authors: Olaf Beckonert, Mary Bollard, Tim Ebbels, Hector Keun, Henrik Antti, Elaine Holmes, John Lindon, Jeremy Nicholson
Biological Chemistry, Biomedical Sciences Division; Imperial College of Science, Technology & Medicine; Sir Alexander Fleming Building, Exhibition Road; South Kensington, London SW7 2AZ; United Kingdom

Keywords: PCA, HCA, KNN, metabonomics, NMR, toxicity classification

Presenter: Olaf Beckonert, o.beckonert@ic.ac.uk

The COnsortium for MEtabonomic Toxicology (COMET) project aims to construct databases and metabolic models of drug toxicity using ca. 100,000 ^{1}H-NMR spectra of biofluids from animals treated with model compounds. Mathematical models characterizing the effects of toxins on endogenous metabolite profiles will enable rapid toxicological screening of potential drug candidates and discovery of novel mechanisms and biomarkers of specific types of toxicity.

In a first attempt to explore and predict the toxicity of 18 model compounds using NMR-metabonomic analysis of biofluids, the individual toxins were administered in separate studies at a toxic (high) and sub-toxic (low) dose. Urine samples were collected from dosed and control animals at 10 time points over a period of 8 days and were subsequently analyzed by 600 MHz ^{1}H-NMR spectroscopy. In order to predict toxicity and to reveal similarities in the response of animals to different toxins, principal component analysis (PCA), hierarchical cluster analysis (HCA) and k-nearest-neighbor (KNN) classification were applied to the data from the high-dose studies to reveal dose and time-related effects. Both PCA and HCA provided valuable overviews of the data, highlighting characteristic metabolic perturbations in the urine spectra between the 3 groups of toxins, namely pancreatic, liver and kidney, and revealed further differences between subgroups of liver toxins. KNN analysis of the multivariate data using both leave-one-out cross-validation and training/test set (50% vs 50%) classification successfully predicted the different toxin classes. In a study-by-study comparison 81% of the samples were predicted into the correct toxin study (50/50 training/test). Classifying the data at a much more detailed level using the individual time points as classes resulted in a success rate of 64%.

This work illustrates the high power and reliability of metabonomic data analysis using ^{1}H-NMR spectroscopy together with chemometric techniques for the exploration and prediction of toxic effects.

Title: Transfer of calibration for classification problems

Authors: Marlana Blackburn, Scott Ramos, Brian Rohrback
Infometrix, Inc.; PO Box 1528; Woodinville, WA 98072; USA

Keywords: calibration transfer, qualitative analysis, Procrustes

Presenter: Marlana Blackburn, marlana_blackburn@infometrix.com

Regression transfer of calibration problems are well-studied. Similar difficulties may arise when a multivariate classification model is created with independent variable data from one instrument and applied to data collected on another instrument. Minor between-instrument differences (*e.g.*, in chromatographic stationary phase, mass spectral tuning, or wavelength registration) can significantly reduce model reliability. When it is impractical to rebuild the model on the second instrument, instrument differences can be mitigated by adjusting profiles from one instrument to look like those collected on the other, using Procrustes analysis. It is critical to include, at a minimum, the same number of transfer samples per class as that class intrinsic dimensionally. Transferability is assessed for qualitative models (KNN and SIMCA) using examples from several instruments.

Title: Mutual peak resolution and matching in series of HPLC/DAD mixture analyses

Authors: Andrey Bogomolov[1], Michael McBrien[2]

1 Advanced Chemistry Development, Inc.; 6 Akademika Bakuleva; 117513 Moscow; Russia

2 90 Adelaide Street West; Toronto, Ontario M5H 3V9; Canada

Keywords: peak matching, multivariate analysis, self-modeling curve resolution, HPLC

Presenter: Andrey Bogomolov, bogomol@acdlabs.ru

One of the largest challenges in chromatographic method development is the necessity for resolving and tracking the movement of peaks as the chromatographic method is changed. A method for mutual peak matching in a series of HPLC/DAD analyses of the same unknown mixture acquired at varying separation conditions has been developed. The approach does not require any prior knowledge of the mixture composition. Applying multivariate analysis techniques to the data matrices the algorithm detects the number of mixture components and calculates retention times of every individual compound in each of the input chromatograms. A common set of UV-Vis spectra of pure components can be obtained as well as their corresponding concentration profiles in separate runs providing full resolution of overlapped peaks. Quantitative criteria of successful peak matching are proposed.

The algorithms were programmed in MATLAB and tested on a number of simulated and real experimental data. Different self-modeling algorithms were compared in order to improve the method performance in the curve resolution stage. Possible ways to improve stability of results, reduce calculation time and minimize operator interaction are discussed. The technique can be used to optimize HPLC analysis of a complex mixture without preliminary identification of its components.

Title: A real-time approach to factor resolution and data refinement using WEFA

Authors: Tara Bohinc[1], Brian Dable[1], Karl S. Booksh[1], Brian Marquardt[2]

1 Arizona State University; PO Box 1604; Phoenix, AZ 85287-1604; USA

2 CPAC; University of Washington; Box 351700; Seattle 98195-1700; USA

Keywords: EFA, trilinear data, unimodality, correlation coefficient convergence

Presenter: Tara Bohinc, Tara.Bohinc@asu.edu

We are investigating a real-time approach to spectral and chromatographic image profiling of trilinear data sets acquired from chemical reaction processes via continuous liquid-chromatographic analysis with hyphenated-instrument detection. The method is based on a unique Window Evolving Factor Analysis algorithm that achieves factor resolution through a step-wise, moving-window technique, while simultaneously allowing data refinement. Pre-processing constraints include unimodality in the chromatographic and spectral directions, in addition to non-negativity applied uniquely in the spectral direction. Model validation is achieved by correlation coefficient guided convergence and is tested on models with differing numbers of factors.

Title: Process monitoring and control in GSK

Authors: Phil Borman, Richard Escott, Jason Murtagh, John O'Shea, Sarah Stimpson
GlaxoSmithKline; Strategic Technologies, Chemical Development;
Gunnels Wood Road; Stevenage SG1 2NY; United Kingdom

Keywords: multivariate SPC, process monitoring, control

Presenter: Phil Borman, pjb79400@gsk.com

A variety of MSPC techniques have been applied on processes in GlaxoSmithKline. Important variables associated with impurity formation in chemical processes have been identified using novel data analysis methodologies such as: data pre-processing algorithms to filter out, or interpolate between, missing values; Multivariate Dynamic Time Warping to achieve batch length equalization; Data Augmentation to enhance the size of the data set in order to extract more information from the limited data and Multi-way Principal Component Analysis (MPCA). The application of MPCA has enabled an overview of batch characteristics to be obtained.

Through initial analyses such as the above the advantages of MSPC has become clear. Thus the next logical step in the application of the MSPC techniques has been to assess their on-line potential. Data from the above chemical processes together with data from GSK fermentation processes, are being used in benchmark studies. The on-line system provides a fingerprint of the process that can be summarized in one or two multivariate control charts. The technique also provides the opportunity for models to include quality parameters (*e.g.*, yield/purity) that allow predictions to be made in real time of the quality of the final product.

Title: The use of on-line near-IR and PLS models for monitoring continuous reactions

Authors: Duncan Thompson, Richard Escott
GlaxoSmithKline; Strategic Technologies, Chemical Development;
Gunnels Wood Road; Stevenage SG1 2NY; United Kingdom

Keywords: NIR, PLS models, continuous reactions

Presenter: Phil Borman, pjb79400@gsk.com

On-line analysis is becoming increasingly important to generate real time quantitative data to monitor the progress of chemical reactions and processes. Spectroscopic techniques can be employed for monitoring continuous processes and the data used to measure achievement and maintenance of steady state equilibrium. The data can also be used to control the quality of product collected after that time point.

This poster describes the use of on-line NIR with an adjustable pathlength flowcell that has been used for monitoring continuous reactions. Chemometric approaches using PCA (Principal Component Analysis) and PLS (Partial Least Squares) have been used to interrogate the data and to build training sets in combination with HPLC assay values. These models have been used to predict product concentrations with further NIR data sets collected during subsequent continuous hydrogenation runs, and the models validated by various techniques.

Title: Application of multivariate curve resolution for analysis of FT-IR microspectroscopic images of *in situ* plant tissue

Authors: Boiana Budevska
DuPont Crop Protection; StinEHaskell Research Center; PO Box 30;
Newark, DE 19714-0030; USA

Keywords: FT-IR hyperspectral imaging, MCR, corn tissue

Presenter: Boiana Budevska, Boiana.O.Budevska@usa.dupont.com

The chemometric techniques of multivariate curve resolution (MCR) are aimed at extracting the spectra and concentrations of individual components present in mixtures using a minimum set of initial assumptions. Results from the application of an alternating least squares (ALS) based MCR to the analysis of hyperspectral images of *in situ* biological material will be presented.

The spectra of individual pure components were mathematically extracted and then identified by searching the spectra against a commercial library. No prior information about the chemical composition of the material was used in the data analysis. The spectra recovered by ALS-MCR analysis of a FT-IR micro-spectroscopic image of 8 micron corn kernel section matched very well the spectra of the corn storage protein, zein, and starch. Through the application of MCR, we were able to show the presence of a second spectrally different protein, that could not be easily seen using univariate analysis. Additional resolution improvements were achieved through application of spectral and concentration constraints together with the common non-negativity constraints. These results demonstrate the value of multivariate curve resolution techniques for the analysis of biological tissue. The value of principal components analysis (PCA) for hyperspectral image analysis is also discussed.

Title: Multi-resolution analysis for quantification of optical properties in scattering media using photon time-of-flight measurements

Authors: C.E.W. Gributs, David H. Burns
Department of Chemistry; McGill University; 801 Sherbrooke Street W;
Montreal, Quebec H3A 2K6; Canada

Keywords: fractal, chaotic, Haar, wavelet, photon time-of-flight

Presenter: David H. Burns, david.burns@mcgill.ca

A multiresolution chaotic systems analysis approach for independent quantification of absorption and scattering properties of scattering samples from time-resolved photon distributions is described. A series of photon time-of-flight measurements were acquired from known composition liquid and granular samples. An analysis of the correlation dimension determined using the time profiles identified low-frequency temporal regions that showed fractal character. The slope of the low frequency component was affected by changes in scattering or absorption. To extract the optical properties, Haar transform based multi-resolution analysis of the experimental time-resolved data was made. The binary character of the Haar transform has been shown to be useful in analyzing fractal signals. Wavelets resulting in the best estimate of the optical properties were selected using the genetic algorithm (GA) and stepwise multilinear regression (SMLR). Using independent data sets, the standard errors of the estimations were significantly lower than in traditional analysis approaches. Since Haar transforms can be implemented using switches, these results indicate that inexpensive instruments may be developed that would allow for a robust quantification in highly scattering media.

Title: Uninformative variable selection using two-dimensional correlation analysis

Authors: He Xiao, David H. Burns
Department of Chemistry; McGill University; 801 Sherbrooke Street W;
Montreal, Quebec H3A 2K6; Canada

Keywords: variable selection, calibration

Presenter: David H. Burns, david.burns@mcgill.ca

A method for the selection of variables in multivariate data sets based on two-dimensional spectroscopy is presented. The method uses random noise equivalent to the level of instrumental noise that is appended to the measured spectra. Two-dimensional correlation spectra of the new matrix are then determined. A series of criteria for variable selection based on the resulting synchronous or asynchronous spectra is shown. Partial-least-square (PLS) analysis is followed on selected variables. Results using both simulated data and measured near-infrared spectra show that a lower standard error of the estimate can be achieved as compared to the use of the entire spectra.

The variables selected are compared to results using other selection methods, such as uninformative variable elimination-PLS, and stepwise multi-linear regression. The regression results for the methods are shown to be comparable. In addition, the sensitivity of the variables selected to noise in the spectra was determined using random noise added in the measurement. The two-dimensional correlation approach is shown to be less sensitive to noise in the spectra as compared to the other methods. The results suggest that selection based two-dimensional correlation analysis may provide a useful tool for rapid determination of variables in a calibration.

Title: A regression-based prediction combining clustering and variable selection methods

Authors: Il-Gyo Chong, Chi-Hyuck Jun
POSTECH; Department of Industrial Engineering; Pohang University of Science and Technology; San 31 Hyoja-dong; Pohang 790-784; Republic of Korea

Keywords: regression, clustering, variable selection, many predictors, global model, local model

Presenter: Il-Gyo Chong, chig@postech.ac.kr

Regression model provides an adequate and interpretable description of how the predictors affect the response. However, in case of large number of predictors, regression model often suffers from its poor prediction accuracy and interpretation. To overcome these drawbacks, many statisticians propose various methods: variable subset selection (best subset regression, forward stepwise selection, backward stepwise regression, stepwise regression, etc.), coefficient shrinkage (ridge regression, the Lasso, etc.), and methods using derived input directions (PCR, PLS, etc.). Variable subset selection has good points of producing an easy interpretable model by retaining a subset of the predictors and discarding the rest while the others have difficulty in interpretation because they predict responses with linear combinations of original predictors.

Clustering method is a descriptive task that seeks to identify homogeneous groups of observations based on the values of their predictors. Observations within each group are more closely related to one another than observations assigned to different groups. In case of large number of predictors, clustering method can be used to reduce the dimensions of predictors by grouping predictors having similar properties. This can be particularly useful when large number of observations is available such as in process industrial.

This paper proposes a prediction method combining clustering method and stepwise regression. This method divides observations of predictors into some local groups having homogeneous observations over some predictors and heterogeneous observations over the others. The former are considered as global predictors explaining global behavior of a response and the latter as local predictors explaining local behavior. Both are independently used to fit global and local model by stepwise regression. Eventually, both global and local models will predict a response given a new observation of predictors. This paper also compares the performance of the proposed method with stepwise regression via a real data example obtained in a steel process. A real example showed that the proposed method results in 16.27% decrease of estimated prediction error compared with the typical stepwise regression.

Title: A response surface study to attain accuracy in measuring atmospheric black carbon by thermal-optical analysis

Authors: Joseph M. Conny, Donna B. Klinedinst
Surface Science and Microanalysis Division; National Institute of Standards and Technology; 100 Bureau Drive, Stop 8372; Gaithersburg, MD 20899-8372; USA

Keywords: response surface modeling, central-composite factorial design, particulate black carbon, atmospheric elemental carbon

Presenter: Joseph Conny, joseph.conny@nist.gov

A product of incomplete combustion, black carbon is a ubiquitous and chemically complex component of atmospheric particulate matter, appearing at measurable levels in even the most remote locations. Black carbon in fine particulate matter (diameter <2.5 microns) has been a major concern for many years because it harmfully impacts health, perturbs climate, and reduces visibility. Accuracy in black carbon measurement has been an elusive goal because of a fundamental conundrum: for such a chemically complex substance, optical properties alone cannot quantify both black carbon mass and the varying aerosol extinction coefficient upon which mass measurement relies. Thermal-optical analysis (TOA), a method of measuring thermally-desorbed carbon from particles while monitoring them optically, has been used to overcome the measurement conundrum because it requires no knowledge of absorptivity. However, variation in the TOA temperature program produces widely varying results, and a consensus on the temperature program is lacking.

TOA involves several thermal desorption steps from 120 ¡C to 900 ¡C of varying duration in either an inert or oxidizing gas stream, resulting in numerous factors that potentially affect accuracy. As a way of optimizing instrument conditions for accuracy, we modeled the TOA response surface for three types of samples: indoor laboratory air, outdoor urban air, and forest fire emissions. Key to TOA is removal of char produced during the higher-temperature steps, which is chemically similar to the native black carbon. We employed a four-factor central composite design to assess variability in the ratio of black carbon to total carbon (BC/TC) and charring. Response surfaces for BC/TC and laser transmission through the sample during charring were modeled with a full second-order polynomial containing 15 terms for main effects, self-interactions, and cross-factor interactions. Optimal conditions for accuracy were revealed from the intersection between two surfaces: the overall minimum laser response and the laser response at the end of the method s charring region.

Title: Optisim: A fast alternative to the Kennard and Stone algorithm

Authors: Michal Daszykowski, B. Walczak, D.L. Massart
Farmaceutische en Biomedische Analyse; Vrije Universiteit Brussel;
Laarbeeklaan 103; B1090 Brussels; Belgium

Keywords: uniform design, OptiSim, representative subset

Presenter: Michal Daszykowski, mdaszyk@vub.ac.be

Very often, data sets are too large to deal with. Usually in this case, a subset of representative objects is selected for further study. The final conclusions about the data set are the generalization of the results obtained for a subset. In many areas, the selection of the representative objects is of great importance, for instance in data mining, drug design, calibration, etc. Uniform subset selection seems to be one of the most often used selection methods. To select a subset uniformly distributed over the data space, the Kennard and Stone algorithm [1] is usually applied. A more recent uniform approach, OptiSim, was proposed by Clark [2]. The fundamental concept explored in OptiSim is to work with random subsamples of S objects, which facilitates and speeds up the computations. The subsample is constructed in an iterative way, and at each iteration S objects, which fulfill some conditions are selected to the subsample. Contrary to the Kennard and Stone algorithm, where at each iteration one out of all data set objects is considered as potential member of the subset, in OptiSim only one object from a subsample ($S<<m$), is added to the subset. The main advantage of OptiSim is its computational efficiency compared to the Kennard and Stone algorithm. Hence, OptiSim is recommended when the data set itself is large. The performance of OptiSim and the Kennard and Stone algorithm on different data sets is illustrated and for both algorithms the computational properties are pointed out.

1) R.W. Kennard, L.A. Stone. Computer aided design of experiments, Technometrics 1969;11:137-148.
2) R.D. Clark. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets, J Chem Inf Comput Sci 1997;37:1181-1188.

Title: Discovering data topology with Growing Neural Gas

Authors: Michal Daszykowski, B. Walczak, D.L. Massart
Farmaceutische en Biomedische Analyse; Vrije Universiteit Brussel;
Laarbeeklaan 103; B1090 Brussels; Belgium

Keywords: Growing Neural Gas, clustering, data mining

Presenter: Michal Daszykowski, mdaszyk@vub.ac.be

The Growing Neural GAs (GNG), proposed by Fritzke [1] aims to describe data topology. It is a neural network, consisting of k nodes, connected by edges. In the training stage the nodes are redistributed over the experimental space to represent the data topology. The GNG network has a dynamic character, since during training new nodes and edges are inserted or removed from the network according to rules, that can be easily modified [2]. The network structure is changed till a stopping criterion is reached. Usually, as a stopping criterion a maximal number of nodes introduced into the network is used. However different criteria can be applied. The most attractive properties of GNG are speed and guaranteed convergence.

These properties make GNG a very powerful tool in describing data set topology. The GNG can be regarded as a clustering technique, because, once the network is trained, each data set object is assigned to its closest node. Additionally, it can perform a selection.

1) B. Fritzke, A Growing Neural Gas network learns topologies, Advances in Neural Information Proceedings Systems 7, MIT Press, Cambridge MA, 1995

2) B. Fritzke, Be busy and unique or be history, the utility criterion for removing units in Self-Organizing networks, KI-99: Advances in Artificial Intelligence 1999;1701:207-218.

Title: Novel combination of hard- and soft-modeling for equilibrium systems and its application to the quantitative analysis of pH modulated mixture samples

Authors: Josef Diewok[1], Anna de Juan[y], Marcel Maeder[y], Rom Tauler[y] , Bernhard Lendl[y]

1 Institute of Chemical Technologies and Analytics; Vienna University of Technology; Getreidemarkt 9/164; A1060 Vienna; Austria

2 Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

3 Department of Chemistry; University of Newcastle; Callaghan, NSW 2308; Australia

Keywords: curve resolution, Newton-Gauss-Marquardt, soft-modeling, hard-modeling, equilibrium, titration, diprotic acids, FT-IR

Presenter: Josef Diewok, jdiewok@mail.zserv.tuwien.ac.at

Second-order calibrations where a spectral data matrix per sample is used instead of a single spectrum are of great interest in analytical chemistry as they allow for the quantification of analytes in presence of unknown and uncalibrated interferents and do not require large calibration data sets as *e.g.* in PLS. In a previous work [1] we presented automatic FT-IR flow titrations of aqueous mixture samples of diprotic organic acid analytes with and without a sugar interferent. Accurate quantitative determination of the acid contents in the mixtures by second order calibration was possible by calculating second derivative data for elimination of baseline contributions and simultaneous analysis of mixture and acid standard titrations with multivariate curve resolution — alternating least squares (MCR-ALS).

However, second order calibrations could not be established if only one diprotic acid was regarded as analyte and the other as an unknown interferent. This is a result of the complexity of the data sets: The pK_a values of the two diprotic acids studied (malic and tartaric acid) are very similar and thus the concentration profiles of the different acid species too correlated for successful resolution of the mixture systems.

In the present contribution we describe how the above-mentioned difficulties in data analysis have been overcome. The pH was measured during the titration of the samples and included in a novel data analysis routine that combines soft- and hard-modeling features. A hard pH equilibrium model was implemented as an additional constraint in the soft MCR-ALS program. This approach allows the successful resolution and quantitation of the analyte diprotic acid, which is subjected to the equilibrium constraint, whereas the interferent diprotic acid is purely soft-modeled. The new algorithm and its application to the analytical problem are discussed in detail. Due to the flexible implementation of the hard-model constraint this new approach is expected to be useful also for analysis of other equilibrium based chemical systems.

1) J. Diewok, A. Juan, R. Tauler, B. Lendl. Quantitation of Mixtures of Diprotic Organic Acids by FTIR Flow Titrations and Multivariate Curve Resolution. Appl Spectrosc 2002;56:40-50.

Title: Modeling the adsorption of carboxylic acid vapors on a quartz crystal microbalance coated with polyethylenimine

Authors: Martha E. Dominguez[1], R.L. Curiale[2], S.M. Steinberg[3], E.J. Poziomek[2], A. Quere[1]

1 Division de Estudios de Posgrado; Facultad de Quimica; Universidad Nacional Autonoma de Mexico; Ciudad Universitaria; Mexico City; Mexico

2 Harry Reid Center for Environmental Studies; University of Nevada Las Vegas; 4505 Maryland Parkway; Las Vegas, NV 89154; USA

3 Chemistry Department; University of Nevada Las Vegas; 4505 Maryland Parkway; Las Vegas, NV 89154; USA

Keywords: carboxylic acids, polyethylenimine, quartz crystal microbalance, mass sensors, adsorption, isotherm, modeling adsorption

Presenter: Martha Dominguez, rolivero@bellsouth.net

Understanding how vapors interact with polymeric coatings and being able to model their adsorption behavior can help in the design of selective and sensitive coatings for use in mass sensors. Many analytes of interest for the application of sensor technology are of acidic nature, thus there is a need to elucidate how coatings react to acidic organic compounds.

In this work, the interaction between a homologous series of carboxylic acids and a basic coating (polyethylenimine) was studied by observing the effect of the increase of the carbon chain in the adsorption at different concentrations of the vapor analytes. A RSM (response surface method) was found to be very useful for simultaneously modeling the effect of the chain length and vapor concentration on the adsorption behavior of carboxylic acids. Published absorption models take in consideration parameters such as hydrogen bonding, acidity and basicity, polarizability, dipolarity, and dispersion forces to identify and characterize the adsorption. The RSM model indicates that the driving parameter on the mechanism of adsorption for carboxylic acids depends on the concentration of the vapor.

Carboxilic acid adsorption on a basic coating is expected to be dominated by a strong acid-base interaction, which was confirmed at low concentration, where the carboxylic acids (methanoic, ethanoic, propanoic and butanoic) adsorbed at about the same concentration. However, at higher concentrations the carboxylic acids with longer carbon chain are adsorbed at a higher concentration than the shorter ones. This is due to the increasing effect of a second important interaction, van der Waals forces, where the molecules of vapor increase interaction between themselves as the concentration of the vapor increases and molecules are considered to start condensing on the surface of the mass sensor. This multi-analyte modeling technique is a viable extension to adsorption curves of BET form that have focused on the behavior of one vapor at the time to enhance the understanding of discriminant factors in the adsorption of compounds with similar functional groups.

A RSM (response surface method) was found to be very useful simultaneously modeling the effect of carboxylic chain length and vapor concentration on the adsorption behavior and finding the interaction between these two factors. In this study the set of vapors was modeled successfully with an R^2 of 0.90.

Title: Rare-earth glass reference materials for near-IR spectrometry

Authors: David L. Duewer, Steven J. Choquette
Analytical Chemistry Division; National Institute of Standards and Technology; 100 Bureau Drive Stop 8394; Gaithersburg, MD 20899-8394; USA

Keywords: PCA, material homogeneity, optical filters, spectrometer x-axis calibration, temperature correction

Presenter: David Duewer, david.duewer@nist.gov

The National Institute of Standards and Technology recently introduced two rare-earth glass optical filter standards for the X-axis (wavelength/wavenumber) calibration of near-infrared (NIR) spectrometers operating in transmittance mode [1,2]. A similar standard designed for diffuse-reflectance mode will be introduced in early 2003. These materials are primarily intended for spectrometer X-axis verification and calibration, but have proven useful for diagnosing subtle instrumental flaws and instabilities as well as being quite good thermometers. They may also have application to some calibration transfer methods.

This poster will summarize our use of low dimensional multivariate analysis techniques required for the Certification of seven NIR band locations in these materials:

1) Material heterogeneity. Principal component analysis enabled the identification and isolation of several environmental and spectrometer sources of x-axis location variability including temperature, humidity and power fluctuations. There proved to be virtually no variability due to material composition differences among the 85 filters evaluated.
2) Instrument performance changes. Knowledge of temperature vs band shift relationships simple error propagation and exploratory graphical analysis enabled isolation of systematic bias between two spectrometers used in the Certification process. Similarly simultaneous analysis of all seven bands enabled identification of within-spectrophotometer performance changes over time and between measurement protocols.

1) SRM 2035 Certificate, National Institute of Standards and Technology, Standard Reference Material 2035 Near Infrared Transmission Wavelength Standard from 10300 cm^{-1} to 5130 cm^{-1} ,Standard Reference Materials Program, NIST, Gaithersburg, MD 20899, 22 February 1999. http://patapsco.nist.gov/srmcatalog/common/view_detail.cfm?srm=2035
2) SRM 2065 Certificate, National Institute of Standards and Technology, Standard Reference Material 2065 Ultraviolet-Visible-Near-Infrared Transmission Wavelength/Vacuum Standard , Standard Reference Materials Program, NIST, Gaithersburg, MD 20899, 28 March 2002. http://patapsco.nist.gov/srmcatalog/common/view_detail.cfm?srm=2065

Title: Imaging spectroscopy: A challenge for multivariate curve resolution methods

Authors: Ludovic Duponchel, W. Elmi-Rayaleh, C. Ruckebusch, J-P. Huvenne, P. Legrand
Universite de Lille; LASIR, Bat C5; F59655 Villeneuve d'Ascq; France

Keywords: imaging spectroscopy, MCR, SIMPLISMA, MCR-ALS, OPA

Presenter: Ludovic Duponchel, ludovic.duponchel@univ-lille1.fr

In analytical chemistry, large matrices are collected from sets of samples under different measurement conditions. These data matrices can be decomposed into a product of two other matrices with a physical or chemical meaning. Many methods have been developed for the resolution of two-way data. Thus the application of multivariate curve resolution methods to electrochemistry, chromatography or spectroscopy makes it possible to retrieve many profiles like spectra, elution profile, pH profile, concentration, and time profile of several components in unresolved and unknown mixtures. The main advantage of these resolution methods is that no prior information about the nature and the composition of analyzed mixtures is necessary.

In imaging spectroscopy, a systematic spectral analysis (*e.g.*, spectral mapping) is carried out with a fixed step size shift over a large sample area. As spectroscopic images are currently obtained by the integration of spectral ranges belonging to only one component, it is very important to develop new methodologies for imaging spectroscopy to extract real contributions. The previous integration method implies the knowledge of all the compounds present in the analyzed sample as well as their specific spectral ranges in order to obtain the corresponding images. These requirements are seldom fulfilled when complex samples of natural or industrial origins are analyzed. Applying multivariate curve resolution methods to imaging spectroscopy datasets permit us to retrieve spatial distributions (images) and spectral signatures of components present in the analyzed sample with no prior knowledge.

In the presented work, the results of a comparative study on some different curve resolution algorithms applied to imaging spectroscopy are discussed. Hence, methods like SIMPLISMA, MCR-ALS, OPA, and PCA are studied alone or combined. Synthetic spectral data are used in order to evaluate the image resolution accuracy. The influence of the spectral bandwidth is studied thanks to the use of mid- and near-infrared reference spectra. Moreover, in order to simulate hard spectral analysis conditions, we study the influence of the signal/noise ratio on image resolution. For this purpose, we calculate a dissimilarity index to compare extraction results based on the correlation between the extracted data (spectra and images) and the reference data (synthetic ones). According to these experiments the combination of OPA and ALS shows a promising resolution ability retrieving the significant pure spectra and the pure image even for low quality spectral data.

Title: On-line determination of polymer (HDPE) properties by low-density nuclear magnetic resonance

Authors: Alan D. Eastman[1], Ping-Chia Liao[2]

1 Phillips Petroleum Co.; 152 Petroleum Lab; Bartlesville, OK 74004; USA

2 Chevron Phillips Chemical Co., LP

Keywords: polyethylene, NMR, density, melt index

Presenter: Alan Eastman, adeastm@ppco.com

Utilizing the free-induction decay curves obtained from a low-field (20 MHz) NMR spectrometer and chemometric analysis, it has been demonstrated that several useful polymer properties can be predicted on line. The analyzer is currently operating in a commercial high-density polyethylene plant to predict polymer density and melt index (*i.e.*, melt viscosity) for over 30 different resin grades.

This paper describes development of the chemometric models, as well as how those models were integrated into the instrument vendor s hardware/software package — which was designed for the vendor s own software. The system is used to make real-time decisions on plant operation, and is being considered for installation at several other HDPE plants worldwide.

Title: Toxicity classification from metabonomic data using a density superposition approach: CLOUDS

Authors: Tim Ebbels, Hector Keun, Mary Bollard, Henrik Antti, Olaf Beckonert, Elaine Holmes, John Lindon, Jeremy Nicholson
Biological Chemistry, Biomedical Sciences Division; Sir Alexander Fleming Building; Imperial College of Science, Technology & Medicine;
South Kensington, London SW7 2AZ; United Kingdom

Keywords: metabonomics, probabilistic NNs, toxicity classification

Presenter: Tim Ebbels, t.ebbels@ic.ac.uk

Predicting the likely toxicity of candidate drugs is of fundamental importance to the pharmaceutical industry. The Consortium for Metabonomic Toxicology (COMET) project aims to construct databases and metabolic models of drug toxicity using ca. 100,000 ^{1}H NMR spectra of biofluids from laboratory rats and mice treated with model compounds. Chemometric methods are being used to characterise the time-related and dose-specific effects of toxins on the endogenous metabolite profiles. Here we present a probabilistic approach to the classification of a large data set of COMET samples using CLassification Of Unknowns by Density Superposition (CLOUDS), a non-neural classification technique developed from probabilistic neural networks.

NMR spectra of urine from rats from 19 different treatment groups, collected over 8 days, were processed to produce a data matrix with 2840 samples and 205 spectral variables. The spectra were normalised to account for gross concentration differences in the urine and regions corresponding to non-endogenous metabolites (~0.5% of the data) were treated as missing values. An initial analysis using samples corresponding to just one animal study showed that a CLOUDS model built with 50% of the samples could successfully predict the time course of the remaining 50%. Almost all misclassifications could be explained by the considerable overlap between adjacent time point classes, as seen in a parallel PCA analysis. When the full data set was classified according to organ of effect, again with a 50/50 train/test set split, over 90% of the test samples could be classified as belonging to the correct group. In particular, samples from liver and kidney treatments were classified with 77% and 90% success respectively, with only a 2% misclassification rate between these classes. Further modeling of the data, counting each of the 19 treatment groups as separate classes, resulted in a mean success rate across groups of 74%. Finally, as a severe test, the data were split into 88 classes, each representing a particular toxin at a particular time point and again 50% of the data removed to use as a test set. 43% of the spectra from non-control samples were classified correctly on a first pass analysis, particularly successful when compared to the null success rate of ~1% expected from random class assignment. The CLOUDS technique has advantages when modeling complex multi-dimensional distributions, giving a probabilistic rather than absolute class description of the data and is particularly amenable to inclusion of prior knowledge such as uncertainties in the data descriptors.

This work shows that it is possible to construct viable and informative models of metabonomic data using the CLOUDS methodology, delineating the whole time course of toxicity. These models will be useful in building hybrid expert systems for predicting toxicology, which are the ultimate goal of the COMET project.

Title: Multiway calibration for creatinine determination by the Jaff method in human serum

Authors: M.V. Guterres, Marcia M.C. Ferreira, P.L.O. Volpe
Instituto de Qu mica - UNICAMP; Universidade Estadual de Campinas; Campinas, S o Paulo 13081-970; Brazil

Keywords: creatinine, PARAFAC, Jaff method

Presenter: Marcia M. C. Ferreira , marcia@iqm.unicamp.br

The creatinine level in human serum samples is currently accepted as an indication of the presence or absence of renal failure, because the concentration of creatinine in serum is independent of the nutritional diet. Despite the trend toward the use of enzymes to improve selectivity for creatinine determination, the classical Jaff reaction still is used extensively for the analysis of creatinine in serum. The Jaff method is based on the spectrophotometric detection of the complex formed when creatinine reacts with picric acid in an alkaline medium. Although this is a common method of choice, a more complete understanding of this reaction system and its interferents than currently exists is needed.

To improve the performance of the specificity of the Jaff method, kinetic methods are used to reduce the effects of the most common interferents: albumin, glucose, biluribin and acetoacetate. Nevertheless, to optimize the kinetic method for all interferents, a better methodology is necessary.

In this work, multiway calibration is proposed to improve the Jaff method. The methodology proposed is based on the simultaneous acquisition of the visible spectrum (450 - 600 nm) at regular time intervals, during a 5 min period. For quantitative determinations, the PARAFAC method is used. The effect of different experimental conditions and the kinetic behavior are investigated for the reaction of creatinine with picric acid in an alkaline medium in the presence of albumin. The results indicate a pseudo-first order mechanism for the creatinine reaction. Determination of creatinine by the PARAFAC method is possible, even when there is a lack of selectivity due to the presence of interferents. The proposed method provides good results and is suitable for creatinine determination in human serum.

This work is financed by the State of S o Paulo Research Foundation (FAPESP).

Title: Chemometric study on atmospheric pollution sources

Authors: Edilton de S. Barcellos[1], Marlon M. dos Reis[1], M rcia M.C. Ferreira [2]

1 Departamento de Qu mica; Universidade Federal de Vi osa; 36570-000 Vi osa; Brazil

2 Instituto de Qu mica - UNICAMP; Universidade Estadual de Campinas; Campinas, S o Paulo 13081-970; Brazil

Keywords: pollution sources, primary pollutants, Tucker model

Presenter: Marcia M. C. Ferreira, marcia@iqm.unicamp.br

This work introduces a methodology to identify the principal emission pollution source in the Regi o Metropolitana de S o Paulo (RMSP). The analysis covered the primary pollutants CO, NO, NO_2 and CH_4, and the secondary one O_3. The data consists of concentration measurements made by Sanitation Department of the State of S o Paulo (CETESB) every hour throughout the year of 1999 for each compound, in the site of P.D. Pedro II. In order to capture the systematic variations for each compound, the data was firstly submitted to a Principal Component Analysis (PCA), on data arranged as matrices 24 (hours of the day) x 365 (days of the year). On the other hand, to extract simultaneously the daily and weekly systematic variations, the data was rearranged in a multiway structure (24 hours of the day x 7 days a week x 52 weeks of the year) and the Tucker model was applied. In this case, there are four modes representing respectively, the source, the weekly, the seasonal and the pollutants contributions.

The results from PCA analysis revealed the daily emission profile for the pollutants CO, NO, NO_2, CH_4 and O_3. The analysis by the Tucker model showed the daily and weekly profiles for the pollutants. From the analyses it was possible to associate CO, NO and NO_2 with the vehicular traffic emission (primary pollutants). CH_4 was identified as a primary pollutant also, but associated primarily with the emissions from another sources. The O_3 was formed by the primary ones (a secondary pollutant).

This work is financed by the State of S o Paulo Research Foundation (FAPESP) and CAPES-PICDT.

Title: Using MATLAB to graphically visualize a batch process with multi-way PCA

Authors: Jennifer Fouche, James Owen, Rajiv Singh
The MathWorks Inc.; 3 Apple Hill Dr.; Natick, MA 01760; USA

Keywords: multi-way PCA, MATLAB

Presenter: Jennifer Fouche, jfouche@mathworks.com

This poster demonstrates the use of graphical projection methods for visualization of process data. The example of condition monitoring for a silicon batch etching process is selected where multi-dimensional projection of the PCA scores evolving within each batch can be used as a graphical tool for describing the batch status. Specifically, the task of visually characterizing the forecasted batch end conditions on basis of the incomplete record of process data available before the end of the batch is addressed.

Multi-way PCA is used to determine a PCA model for a calibration data set consisting of normative batches. By assuming a Gaussian distribution of process variables, the conditional probability distribution of the multi-way PCA scores is determined at a sequence of time steps during the progression of the batch. By graphically projecting of the confidence regions defined by these conditional distributions, the likelihood of aberrations from normal batch behavior can be detected in advance, and appropriate mid-course corrections can be applied. The poster session will end with a demonstration of other potential uses of this visualization technique in fault detection and condition monitoring applications.

Title: Fusing data from diverse sources to characterize batch reactions

Authors: Paul J. Gemperline[1], Shane Moore[1], Enric Comas[1], R. Russell Rhinehart[2], Karen High[2], Samir Alam[2]

1 East Carolina University; Department of Chemistry; 327 Flanagan; Greenville, NC 27858-4353; USA

2 School of Chemical Engineering; Oklahoma State University; Stillwater, OK 74078; USA

Keywords: self-modeling curve resolution, dynamic modeling, batch process monitoring

Presenter: Paul Gemperline, gemperlinep@mail.ecu.edu

In this presentation we report a technique under development for experimental optimization of batch recipes in real-time. Research software receives *in situ* spectroscopic measurements and process measurements from a laboratory batch micro-reactor. Using dynamic modeling, the data from two diverse sources (spectroscopy and calorimetry) is combined into a unified model of spectral profiles, reaction rates, mass balance and energy balance.

Initial experiments based on a batch titration mode of operation are reported, where small aliquots of reagents are delivered to the reactor over a period of time. By monitoring the reactor's time-dependent response after a few small additions, dynamic models are adjusted off-line. In future work, the adjusted models will be used on-line to forecast the location of the batch endpoint. Large reagent additions can then be safely made to reach the endpoint rapidly, thereby compensating for variation in the quality of starting materials.

Title: Monitoring batch processes using OPA and ALS

Authors: S. Gourvenec, D.L. Massart
Farmaceutische en Biomedische Analyse; Vrije Universiteit Brussel;
Laarbeeklaan 103; B1090 Brussels; Belgium

Keywords: Orthogonal Projection Approach (OPA), Curve Resolution, batch, process control, on-line monitoring

Presenter: Sebastien Gourvenec, sgourven@vub.ac.be

Batch processes, which are often characterized by a reaction (or a succession of reactions) between materials that are charged in predefined proportions in a reactor and react for a finite duration, play an important role in the production of high added value products. Since there is a real need to control the process and to detect as soon as possible if the batch is going in a wrong direction to save costs and time, several methods were proposed to monitor batch processes.

The Orthogonal Projection Approach (OPA) applied on spectroscopic data (*e.g.*, near-infrared) recorded during the process can provide a lot of information within the chemical system such as the concentration of species present in the reaction, changes in solvent conditions, presence of impurities, etc. OPA is one of the many different curve resolution methods and resolves the data matrix into the concentration profiles and the spectra of the components present during reaction(s). Working without prior information about the shape of pure spectra and/or concentration profiles, OPA is a self-modeling method that compares the data matrix with a reference and is suitable for on-line process monitoring. Based on the Gram-Schmidt orthogonalization and on the assumption that the purest spectra in the data matrix are mutually more dissimilar than the corresponding mixture spectra, OPA coupled with an Alternating Least Squares (ALS) procedure is presented as a way to monitor batch processes. This method allows predictions of concentration profiles within a batch and these predicted values are useful to follow the evolution of the concentrations of the different species according to time.

An application, taking into account the variations of real batch processes data, to monitor a new batch on-line, is explained and shows OPA as an effective method for batch process analysis.

Title: Parallel column liquid chromatography with a single multi-wavelength absorbance detector for enhanced selectivity using chemometric analysis

Authors: Gwen M. Gross, Bryan J. Prazen, Robert E. Synovec
Department of Chemistry; University of Washington; Box 351700;
Seattle, WA 98195; USA

Keywords: liquid chromatography, parallel column, generalized rank annihilation method

Presenter: Gwen M. Gross, gmlg@u.washington.edu

The selectivity of liquid chromatography separations is increased using a parallel column configuration and chemometric analysis. In this system, an injected sample is first split between two liquid chromatographic columns that provide complementary separations. The effluent from the two columns is recombined prior to detection with a single multi-wavelength absorbance detector. Complementary stationary phases are used so that each chemical component produces a detected chromatographic concentration profile consisting of two peaks.

Although it is counter-intuitive, a parallel column configuration, when coupled with multivariate detection, provides substantially increased chemical selectivity relative to a single column configuration with the same multivariate detection. This enhanced selectivity is achieved by doubling the number of peaks in the chromatographic dimension while keeping the run time constant. Using complementary parallel columns results in a broadly applicable separation system.

Unlike traditional single column separation methodology, the parallel column system sacrifices chromatographic resolution while actually increasing the chemical selectivity, thus allowing chemometric data analysis methods to mathematically resolve the multivariate chromatographic data. The parallel column system can be used to reduce analysis times for partially resolved peaks and simplify initial method development as well as provide a more robust methodology if and when subsequent changes in the sample matrix occur (such as when new interferences show up in subsequent samples).

Here, a mixture of commonly found aromatic compounds were separated with this system and analyzed using the generalized rank annihilation method (GRAM). Analytes that were significantly overlapped on both stationary phases applied, ZirChrom PBD and CARB phases, when used in traditional single column format, were successfully quantitated with a %RSD of around 2% when the same stationary phases were used in the parallel column format. These results indicate that a parallel column system should substantially improve the chemical selectivity and quantitative precision of the analysis relative to a single-column instrument.

Title: Optimization of preparation of plant samples for metabolic profiling by GC-MS

Authors: Jonas Gullberg[1], P. Jonsson[2], A. Nordstr m[1], M. Sj str m[2], M. Kowalczyk[1], G. Sandberg[1], T. Moritz[1]

1 Ume Plant Science Centre (UPSC); Department of Forest Genetics & Plant Physiology; The Swedish University of Agricultural Sciences (SLU); SE901 83 Ume ; Sweden

2 Research Group for Chemometrics; Organic Chemistry; Ume University; SE901 87 Ume ; Sweden

Keywords: metabolic profiling, optimisation, MS, experimental design, PCA, PLS

Presenter: Jonas Gullberg, jonas.gullberg@genfys.slu.se

To be able to extract relevant biological information from datasets obtained by metabolic profiling by mass spectrometry, the metabolic information has to be maximized and at the same time the analytical and biological activity minimized. The aim of this study is to optimize both tissue extraction and derivatization of the metabolome for plant samples using experimental design. Factors to investigate are different sampling treatments and chemical and physical factors during extraction and derivatization.

To understand the variability due to biological activity and analytical errors between the different experimental set-ups, chemometrical methods can be applied on GC-MS data. By using a latent variable method the variation can be described by a small number of independent latent variables. When logical groups occur, supervised classification methods such as PLS-DA or O2-PLS-DA can be used. The scores and loadings show the underlying structures in the data that can be used to interpret differences the between groups of samples.

Title: Analysis of video images used to study gas-liquid transfer

Authors: Stephen P. Gurden, Euler M. Lage, Cristiano G. de Faria, In s Joekes, M rcia M.C. Ferreira
Instituto de Qu mica - UNICAMP; Universidade Estadual de Campinas; Campinas, S o Paulo 13081-970; Brazil

Keywords: chemical imaging, gas-liquid transfer, carbon dioxide exchange, PCA, PARAFAC

Presenter: Stephen Gurden, spgurden@iqm.unicamp.br

The use of chemical imaging is a developing area that has potential benefits for chemical systems where spatial distribution is important. Examples include processes in which homogeneity is critical, such as polymerizations, pharmaceutical power blending and surface catalysis, and dynamic processes such as the study of diffusion rates or the transport of environmental pollutants. The advent of high-resolution spectroscopic imaging instrumentation along with the continual increase in data storage and processing power suggest that chemical imaging will become an important tool in the future.

The exchange of CO_2 between the air and sea is an important process in terms of the global mass cycling system. In this work, we present a study of images taken from an experiment in which the exchange of CO_2 from air to water is investigated under controlled temperature and salinity conditions. The presence of a pH indicator in the water produces a color change that enables the uptake of CO_2 to be followed dynamically using a standard video camera.

Prior to statistical analysis on the data, it is necessary to perform a reconciliation step (image rotation, cropping and resizing) that yields congruent image arrays. The analysis of single multivariate images using the PARAFAC model is then described, and contrasted with PCA, as a form of understanding the relationships between the spatial and wavelength directions. The use of other single image transformations is also described, including the use of histograms (particularly meaningful here) and pH mapping. The analysis of multiple images related in time (*i.e.*, movies) using both PCA and PARAFAC is then described, along with different ways of preprocessing movie arrays. The model components found are used to understand factors common to all the CO_2/water exchange experiments, as well as factors specific to particular experimental runs.

This work is financed by the State of S o Paulo Research Foundation (FAPESP).

Title: A hybrid Genetic Algorithm - Tabu Search approach for optimising multilayer optical coatings

Authors: Jos A. Hageman, R. Wehrens, H.A. van Sprang, L.M.C. Buydens
Laboratory of Analytical Chemistry; Katholieke Universiteit Nijmegen;
Toernooiveld 1; NL6525 ED Nijmegen; The Netherlands

Keywords: optimisation, Genetic Algorithms, Tabu Search, multilayer optical coatings

Presenter: Jos Hageman, hageman@sci.kun.nl

Constructing multilayer optical coatings (MOCs) is a difficult large-scale optimization problem due to the enormous size of the search space. In this poster, a new approach for designing MOCs is presented using Genetic Algorithms (GA's) and Tabu Search (TS). In this approach, it is not necessary to specify how many layers will be present in a design, only a maximum needs to be defined. As it is generally recognized that the existence of specific repeating blocks is beneficial for a design, a specific GA representation of a design is used that promotes the occurrence of repeating blocks. Solutions found by GA's are improved by a new refinement method, based on TS, a global optimization method that is loosely based on artificial intelligence. The improvements are demonstrated by creating a visible transmitting / infrared reflecting filter with a wide variety of materials.

Title: Identification of pure-component spectra from mid-IR spectral data of multi-component mixture using independent component analysis

Authors: Sangjoon Hahn, Haemin Cho, Gilwon Yoon
Medical Application Team; Samsung Advanced Institute of Technology; PO Box 111; Suwon 440-600; Republic of Korea

Keywords: independent component analysis, PCA, mid-IR

Presenter: Sangjoon Hahn, sjhahn@sait.samsung.co.kr

We present a new method that can extract the pure spectra from mid-IR spectra of multi-component mixture using Independent Component Analysis. This is accomplished by making good use of higher-order statistical moments of a spectrum. The main advantage of a new method is that it is able to identify the pure spectra of the constituent components from the spectra of their mixtures without *a priori* knowledge of the mixture. The ICA based method is therefore particularly useful in identifying the unknown components in a mixture as well as in estimating their concentrations.

Examples considered are two-component systems consisting of aqueous solution of glucose and sucrose that exhibit distinct but heavily overlapped spectra. In side-by-side tests using both simulated and experimental data, ICA combined with PCA can identify pure-component spectra as well as the correct number of components in the mixture.

Title: Resolution of humic materials and chlorophylls using light-emitting diode excitation emission matrix (LED-EEM) fluorescence spectroscopy and parallel factor analysis (PARAFAC)

Authors: Renee D. JiJi, Sean J. Hart
Chemistry Division, Code 6116; Naval Research Laboratory; 4555 Overlook Avenue; Washington, DC 20375; USA

Keywords: LED, EEM Fluorescence, PARAFAC, humic material, chlorophyll

Presenter: Sean Hart, shart@ccf.nrl.navy.mil

An excitation emission matrix (EEM) fluorescence instrument has been developed using a linear array of light emitting diodes (LED). The excitation wavelengths covered extend from the upper UV through near infrared (NIR) spectrum: 370 nm to 880 nm. Using a LED array to measure fluorescence emission at multiple excitation wavelengths is a low-cost alternative to a high power lamp and imaging spectrograph. LEDs allow a larger simultaneous excitation range than is possible using a lamp and monochrometer. The array of LEDs is focused into a sample cuvette, creating spatially separated excitation spots. Fluorescence from analytes in solution is collected at a right angle by another lens, which images the fluorescent spots onto the entrance of a spectrograph with a CCD camera for detection. The array is easily tailored to the application by selecting LED wavelengths from the growing commercially available selection.

The fluorescence of humic materials and chlorophyll are commonly used as indicators of dissolved organic carbon and phytoplankton growth in natural waters. In addition, natural waters have characteristic EEM fluorescence spectra based on several factors including origin and bioactivity. Combination of EEM fluorescence spectroscopy and multiway methods allows resolution of individual component excitation and emission profiles and concentration ratios. This information may then be used for classification of each sample and quantification of selected analytes, such as chlorophyll. The utility of this instrument coupled with parallel factor analysis (PARAFAC) has been evaluated using a complex spectroscopic system: humic materials, chlorophylls, and bacteriochlorophyll.

Title: High-speed gas chromatographic separations with diaphragm valve-based injection and chemometric analysis as a gas chromatographic "sensor

Authors: Kevin J. Johnson, Marianne A. Cavelti, Janiece L. Flick, Jay W. Grate, Robert E. Synovec
Department of Chemistry; University of Washington; Box 351700; Seattle, WA 98195; USA

Keywords: gas chromatography,sensor, PLS

Presenter: Janiece L. Hope, janiece@u.washington.edu

A high-speed gas chromatographic system, the gas chromatographic sensor (GCS), is developed and evaluated. The GCS combines fast chromatographic separation and chemometric analysis of the resulting chromatogram in order to produce an instrument capable of high-speed, high-throughput screening and quantitative analysis of complex chemical mixtures on a similar time scale as conventional chemical sensors.

High-speed gas chromatographic separation in the GCS is made possible by the use of short capillary column lengths, high carrier gas flow velocities, and a novel valve-based injection technique that provides significantly narrower injection plugs than standard gas chromatograph (GC) injectors. The system is currently installed inside the oven of a standard gas chromatograph and, for evaluation purposes, utilizes the injector of the gas chromatograph to deliver sample to the diaphragm valve. The resulting chromatograms are on the order of one second in duration, with overall run times for a single sample being somewhat larger due to the time required for the injection system on the GC to provide sample to the valve. It is anticipated that this system will eventually be used to rapidly sample and analyze industrial process streams directly without the use of a standard GC injector.

The GCS was evaluated with test mixtures consisting of fifteen different chemical components from four master mixtures each consisting of one of four different chemical classes: five alkanes, three ketones, four alkyl benzenes, and three alcohols. Twenty-eight different mixture blends were constructed by varying percent volume of each class of compounds. The mixtures were analyzed for percent volume content of two of the four classes of compounds, with the other two classes treated as variable background interference. Each of the 28 samples was subjected to 5 replicate GCS runs. Calibration models to predict percent volume content of alkanes and of ketones were constructed using partial least squares regression on calibration sets consisting of the replicate GCS runs of six different samples. The replicates of the remaining 22 samples were then analyzed for percent volume content. Root mean square errors of prediction were two to three percent of the mean percent volume values for either alkane or ketone prediction models, depending on the GCS column used for analysis and the samples chosen for the calibration set of that model.

Title: Data mining of the relationship between volatile organic components and transient high ozone formation

Authors: Feng Gan[1], Philip K. Hopke[2]

1 Department of Chemical Engineering; Clarkson University; Potsdam, NY 13699; USA

2 Department of Chemistry; Clarkson University; Potsdam, NY 13699-5705; USA

Keywords: data mining, volatile organic components, transient high ozone formation

Presenter: Philip K. Hopke, hopkepk@clarkson.edu

The aim of this research is to identify the relationships between volatile organic components and transient high ozone formation in the Houston area. The ozone is not emitted to the atmosphere directly but is formed by chemical reactions in the atmosphere. In Houston, short-term (1 hour) sharp increases are observed followed by a rapid decrease back to typical concentrations. Automatic gas chromatographs are operated at several sites that cryogenically collect volatile organic compounds (VOCs) during and hour and then the compounds are flash evaporated into the GC for analysis. Chromatographic data for 66 volatile organic components are stored in analysis report text files. A program has been developed to read the amount of each component in the measurements such that a data set is generated that includes the concentrations of each VOC for each hourly sample. A subset of the data is selected that corresponds to the period of the positive ozone transient and these data are used in the data mining process.

Based on chemical mass balance, a linear model was established between the subset and the positive ozone transition. Non-negative least squares was used to calculate the regression coefficient of the volatile organic components that have the most significant positive contribution to the positive ozone transition.

Title: Discarding or down-weighting high-noise variables in factor analytic models

Authors: Philip K. Hopke, Pentti Paatero
Department of Chemistry; Clarkson University; Potsdam, NY 13699-5705; USA

Keywords: FA, positive matrix factorization, PCA, noise

Presenter: Philip K. Hopke, hopkepk@clarkson.edu

It has been reported in earlier factor analysis studies that one may need to exclude noisy variables (concentrations of poorly determined or low abundance elements, say) in source apportionment of environmental particulate matter mass in order to obtain sensible results. The reasons for excluding such variables have not been understood. This work investigates the effect of weak, noisy variables on Principal Component Analysis (PCA) and Positive Matrix Factorization (PMF).

A variable is called weak if the average ratio of signal to noise (S/N) is typically between 0.5 and 2.0. If, however, S/N < 0.5, then the variable is called bad while a variable is good if it's S/N > 2.0. It has been suggested that optimally the variables should be scaled in PCA so that the average errors are the same for all variables in the matrix. It is assumed here that in PCA, all the variables are centered and scaled variable by variable. This work demonstrates that scaling weak or bad variables so that their average error exceeds unity (=upweighting in comparison to standard scaling) is detrimental to the results of PCA or PMF. The random errors in computed factors increase steeply so that the number of resolvable factors is decreased. It is also demonstrated that scaling weak or bad variables to below-unity errors (=downweighting) is sometimes slightly better than standard scaling. Systematic downweighting of weak and bad variables is recommended as an insurance against the occasional upweighting that should be avoided at all costs.

Autoscaling is often used as a preparatory step in PCA. For good variables, autoscaling usually provides a reasonably even scaling so that their average errors are of similar magnitudes. However, autoscaling increases the errors of weak and bad variables above the errors of good variables, and thus prevents an optimal analysis. The main result of this work is that autoscaling should not be applied to weak or bad variables! Instead, the bad variables and their errors should be scaled downward so that their average noise level is typically equal to 10% of the noise level of good variables. Then the average values of bad variables will be a few percent of the average values of good variables. Alternatively, bad variables may be omitted altogether from the model.

To accomplish proper scaling when preparing data for PMF, the error estimates of weak variables should be increased, typically by a factor of three. Bad variables should either be discarded or else their error estimates should be increased by 10, say. Such downweighting will help in resolving weak sources.

Title: Assessment of DOSY data processing methods

Authors: Ruifen Huo[1], R. Wehrens[1], J. van Duynhoven[2], L.M.C. Buydens[1]

1 Laboratory of Analytical Chemistry; Katholieke Universiteit Nijmegen; Toernooiveld 1; NL6525 ED Nijmegen; The Netherlands

2 Central Analytical Science; Unilever Research Vlaardingen; The Netherlands

Keywords: DOSY NMR, diffusion NMR, SPLMOD, CONTIN, MCR, FA, DECRA

Presenter: Ruifen Huo, ruifen@sci.kun.nl

DOSY (Diffusion-Ordered SpectroscopY) NMR is based on pulse-field gradient spin-echo NMR experiment, in which components experience self-diffusion with different diffusion rate as the gradient strength increases, constructing a bilinear NMR data set of a mixture. By calculating the diffusion coefficient for each component, it is possible to obtain a 2D NMR spectrum: one dimension is for the conventional chemical shift and the other for the diffusion coefficient. The most interesting point is that this 2D NMR allows non-invasive chromatography to obtain the pure spectrum for each component, leading to potential substitution for the more expensive and time-consuming LC-NMR. Potential applications of DOSY NMR include identification of the components and impurities in complex mixtures such as body fluids, surfactants, and polymers.

Data processing is the heart of DOSY NMR. Single channel methods and multivariate methods have been proposed for the data processing but all of them have difficulties when applied to real-world cases. The big challenge appears when dealing with more complex samples, *e.g.* components with small difference of diffusion coefficients severely overlapping on the dimension of chemical shift. Two single channel methods including SPLMOD and CONTIN and two multivariate analysis methods called DECRA and MCR are critically evaluated by simulated and real DOSY data sets. The assessments indicate the possible improvement of the DOSY data processing.

Title: Hard-and soft-modeling of acid-base chemical equilibria of biomolecules using ^{1}H-NMR

Authors: Joaquim Jaumot, Montserrat Vives, Raimundo Gargallo, Rom Tauler
Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

Keywords: NMR, hard-modelling, soft-modelling, oligonucleotides, inert-labile, pK values

Presenter: Joaquim Jaumot, joaquim@apolo.qui.ub.es

Application of different spectroscopic techniques allow the direct study and species resolution of chemical equilibrium processes of biomolecules. ^{1}H-NMR spectroscopy is a powerful tool that allow the simultaneous determination of thermodynamic (*i.e.*, pK values) and structural information about species at equilibrium. One of the important considerations to take into account when ^{1}H-NMR spectroscopy is used for studying equilibria is the time scale of measurements and the inert-labile character of the reaction equilibrium under study.

Different situations are encountered depending on the measuring time and exchange equilibrium rates. Acid-base equilibrium reactions are too fast and traditional ^{1}H-NMR measuring times are too slow for an independent differentiation between species signals and only an average signal is available. Therefore traditional approaches for data treatment and equilibrium investigation using ^{1}H-NMR have involved parameter estimation by means of hard-modeling non-linear least squares curve fitting methods. For inert systems however, approaches based in generalized linear models can be used in the same way as for electronic spectroscopies allowing species spectroscopic signals differentiation and resolution.

In this work, these two approaches are compared and used in the study the protonation equilibria of oligonucleotides dGTG, dCMP and dGMP that present strong labile character. A new data pre-treatment is applied to allow the study of a fast labile equilibrium by means of soft-modeling multivariate curve resolution methods requiring inert behavior of measured signals. Comparison of both approaches are shown to give similar pK_a values.

Title: Multivariate curve resolution alternating least squares analysis of the conformational equilibria of the oligonucleotide d<TGCTCGCT>

Authors: Joaquim Jaumot, N ria Escaja, Raimundo Gargallo, Enrique Pedroso, Rom Tauler
Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

Keywords: MCR-ALS, oligonucleotide, conformational equilibria, rank deficiency, rotational ambiguity

Presenter: Joaquim Jaumot, joaquim@apolo.qui.ub.es

DNA conformations different from the classical Watson-Crick double helix have been described in recent years and, in particular, duplex bi-loop structures of d<pTGCTCGCT> are very interesting and relevant because they show how a cyclic sequence may allow self-recognition between DNA molecules.

Equilibria studies between different conformations of polynucleotides and oligonucleotides can be performed by means of melting and/or salt titration experiments monitored using UV molecular absorption or circular dichroism spectroscopies. These experiments have been traditionally performed using a single wavelength approach (univariate data analysis). This approach has clear drawbacks, the most important of them is the assessment of the number of different conformations simultaneously present in one experiment when no selective wavelengths are present. These difficulties can be overcome by multiple wavelength spectroscopy and application of appropriate multivariate chemometric data analysis methods. Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) analysis of individual spectroscopic melting and/or salt titration or experiments may be however, additionally hindered by rank deficiency and rotational ambiguity problems. In these cases, species resolution is not totally achieved, in particular for intermediate conformations present at low concentrations, embedded concentration profiles and with strongly overlapped spectra. To improve these resolution difficulties, multiway MCR-ALS simultaneous analysis of multiple and specially designed melting (temperature changing) and salt titration experiments using two spectroscopic techniques is proposed. This powerful analysis allows identification of different spectroscopically distinct conformations and resolution of their concentration and pure spectra profiles.

In this work, transition equilibria between conformations of the cyclic oligonucleotide d<pTGCTCGCT> upon salt and temperature changes have been studied by means of UV molecular absorption and circular dichroism (CD) spectroscopies and MCR-ALS. Three different conformations were resolved assigned respectively to a random coil conformation, to a dumbbell (monomeric) conformation and to a bi-loop (dimeric) conformation.

Title: Relating physiological measurements to temporal change of sensory attributes in semi-solid food products

Authors: Renger H. Jellema, R.A. de Wijk, L.J. van Gemert, G.B. Dijksterhuis, J.F. Prinz, H. Weenen
TNO Nutrition and Food Research; PO Box 360; Zeist; NL3700 AJ; The Netherlands

Keywords: sensometrics, time, intensity, food, physiology

Presenter: Renger H. Jellema, jellema@voeding.tno.nl

In sensory studies a broad range of techniques are applied to study how food is perceived. Techniques involved in these kind of studies are for instance sensory profiling studies, consumer acceptance, time-intensity experiments (TI), physiological and rheological measurements. The data obtained from this broad range of studies need to be related to each other to understand underlying phenomena that result in the liking or dislike of a set of food products. Armed with this knowledge, product developers are able to improve existing or develop new products that are likely to be accepted by the consumers. In order to obtain the desired relationships, multivariate techniques such as *e.g.* PCA and PLS are used in numerous ways.

In a multidisciplinary project that was initiated in 1999, the above mentioned techniques were applied to uncover the underlying mechanisms involved in the perception of oral texture attributes for semi-solid food products. In this project, physico-chemical parameters were related to perceived texture. First, sensory attributes were generated and used to profile products. Food properties were assessed *in vitro* by a trained sensory panel, rheologists and food chemists. Changes in food properties during mastication were assessed by physiologists in in-mouth experiments. In one of the studies physiological measurements were related to time-intensity measurements. Time intensity measurements are ratings of perceived intensity of a specific sensory attribute, during the time a food product is processed in the mouth. The resulting time-intensity curves are known to contain so-called signature effects which means that subjects from the panel can be recognized from their individual curves. Two possible underlying mechanisms are the translation from perception to the score given by the member of the panel and secondly differences in oral processing between the subjects of the panel. Individual differences include heating, mixing and dilution by saliva, chemical break-down by amylase, and physical break-down by mechanical forces, factors that are known to differ over individuals.

Typical results showing the relationship between results obtained from time-intensity measurements and results from physiological measurements will be presented. These results can be used in future research to get a better understanding of underlying mechanisms. The results may also be used to separate product variation from variation due to differences between subjects from a panel. Initial results of combined time-intensity and physiological measurements indicate that variations between judges exceeded variations between food products, especially in T-I measurements. Also, specific physiological measurements appear to relate to specific sensory measurements (*e.g.*, creamy/sticky) but not to others (*e.g.*, cold).

Title: Non-invasive diagnosis of osteoarthritis using MVA on NMR spectra

Authors: Renger H. Jellema, Robert-Jan A.N. Lamers, Elly J. Faber, Jeroen de Groot, Gerwin K. Spijksma, Nicole Verzijl, Joop H.J. van Nesselrooij
TNO Nutrition and Food Research; PO Box 360; Zeist; NL3700 AJ; The Netherlands

Keywords: NMR, osteoarthritis, non-invasive

Presenter: Renger H. Jellema, jellema@voeding.tno.nl

Osteoarthritis (OA) is one of the most common diseases among the elderly. This common chronic disabling disorder involves the wear and tear on the joints and can cause mild to severe pain and stiffness. Diagnosis and monitoring of OA, especially in an early stage, is difficult. Therefore, the development of sensitive biomarkers for early diagnosis, disease activity and progression is of significant importance. A new method involving non invasive diagnosis by means of Nuclear Magnetic Resonance (NMR) spectroscopy and multivariate techniques (MVA) such as PCDA and PLS was developed.

Biological fluids, such as urine and blood, contain a wealth of information about the metabolic status of a living system, and thus about its clinical or histopathological status. ^{1}H-NMR is a very suitable technique to analyze biological fluids as it provides both quantitative and qualitative information. Nevertheless, ^{1}H-NMR spectra of biological fluids are very complex and variations between them are often too small to be recognized by the eye. Therefore, to find significant differences, MVA methods were used to explore patterns in ^{1}H-NMR data. An important aspect of the data analysis involves the pre-processing of the NMR data. Due to for instance differences in concentrations and biological factors the spectra are not directly comparable to each other. With the combination of pre-processed high resolution ^{1}H-NMR data and MVA it is possible to rapidly analyze biological fluids and thus evaluate the metabolic status.

A study was carried out with guinea pigs developing osteoarthritis during ageing. Clinical and histo-pathological parameters were determined and assessed. PCDA and PLS were carried out on the NMR measurements in combination with these data. Thereupon, a PLS model was constructed, predicting osteoarthritis in guinea pigs. From preliminary experiments it appears that the model is powerful in predicting the histopathological status of unknown samples. Further study is needed to get a better understanding of the relationship between the results of the NMR measurements and OA. The latter can be of significant benefit in the search for new methods of treatment for OA.

Title: Quantification of two-ring aromatics in jet fuel with GCxGC/Tri-PLS and objective retention time alignment

Authors: Kevin J. Johnson, Bryan J. Prazen, Donald C. Young, Robert E. Synovec
Department of Chemistry; University of Washington; Box 351700;
Seattle, WA 98195; USA

Keywords: comprehensive two-dimensional gas chromatography, tri-linear PLS, jet fuel, two-ring aromatic, retention time alignment

Presenter: Kevin Johnson, kjj@u.washington.edu

Tri-linear partial least squares (tri-PLS) is used in conjunction with comprehensive two-dimensional gas chromatography (GC x GC) in order to quantify the percent by volume of two-ring aromatic compounds in jet fuel samples. The increased peak capacity and selectivity of GC x GC makes the technique attractive for the rapid analysis of complex mixtures, such as jet fuel. Analysis of complex mixtures by GC x GC can be further enhanced through the use of second order chemometric techniques such as tri-PLS. Unfortunately, retention time variation has traditionally been an impediment to chemometric analysis of chromatographic data. It is demonstrated here that the effects of run-to-run retention time variation can be mitigated through the application of a retention time alignment algorithm of the entire chromatograms that combines rank minimization and retention time axis stretching with interpolation. A significant decrease in calibration error is observed when the algorithm is applied to chromatograms prior to construction of tri-PLS models.

Fourteen jet fuel samples with known volume percentages of two-ring aromatic compounds (ASTM D1840) were obtained and subjected each to five replicate five-minute GC x GC separations over a period of two days. A subset of seven samples spanning the range of volume percentages of two-ring aromatics was chosen as a calibration set and tri-PLS calibration models were subsequently developed in order to predict the two-ring aromatic compound content of the samples from the GC x GC chromatograms of the remaining samples.

Calibration models constructed from the original, unaligned GC x GC chromatograms exhibited a root mean square error of prediction of five percent of the median volume percent value. Calibration models constructed from GC x GC chromatograms that were retention time corrected, however, exhibited a root mean square error of prediction of three percent of the median volume percent value. The error of prediction is further lowered to one percent of the median volume percent value when the aligned chromatograms are truncated to include only selected regions of the chromatogram populated by large polar molecules. Thus, the combination of feature selected GC x GC coupled with retention time alignment provides excellent accuracy and an optimized prediction error similar to the uncertainty of the known volume percent values.

Title: Applied influence function analysis for partial least squares

Authors: Kjell Johnson, William Rayens
Pfizer Global Research & Development; 2800 Plymouth Road;
Ann Arbor, MI 48105; USA

Keywords: PLS, influence function, empirical influence function, outlier, leverage

Presenter: Kjell Johnson, kjell.johnson@pfizer.com

Since the inception of partial least squares (PLS) with Herman Wold s Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, its use has become widespread. This is partly due to the ability of PLS to obtain a relationship between descriptors and response(s) in the over-determined regression setting. In addition, PLS has been successfully implemented in other regression-type applications such as experimental design and classification, that has bolstered its use. Although PLS can be implemented via the original NIPALS algorithm, it can also be implemented by way of solutions to well-posed eigenstructure problems that emphasize the compromise PLS strikes between summary and purpose.

Unfortunately, as many authors have previously noted, a PLS model can easily be unduly influenced by one or more observations. A common way to identify influential observations in PLS is to use regression diagnostic tools on the scores and the response. While the use of regression diagnostics has intuitive appeal, the tendency to look for outliers and leverage points in the regression sense can be misleading because these tools were not developed to consider the underlying objective of PLS.

As an alternative approach, we have used influence function (IF) theory coupled with the eigenstructure perspective of PLS to develop empirical influence functions (EIFs). These newly derived EIFs enable one to identify observations that directly influence the eigenstructures of PLS, and hence, the performance of the PLS model.

To provide an intuition for the EIFs of both eigenvalues and eigenvectors, we briefly introduce the results of our IF derivations. The primary focus of this work will be to illustrate, through practical examples in chemistry applications, how the EIF is used to identify influential observations. By identifying the appropriate influential observations, one can, in turn, improve the predictive ability of the PLS model.

Title: Geometry-based pattern recognition of dynamic metabolic profiles: Classifying toxicity from NMR spectra of biofluids

Authors: Hector Keun, Tim Ebbels, Henrik Antti , Mary Bollard, Olaf Beckonert, Elaine Holmes, John Lindon, Jeremy Nicholson
Biological Chemistry, Biomedical Sciences Division; Imperial College of Science, Technology & Medicine; Sir Alexander Fleming Building, Exhibition Road; South Kensington, London SW7 2AZ; United Kingdom

Keywords: metabonomics, SIMCA, NMR, toxicity classification, COMET, SMART, geometry-based pattern recognition

Presenter: Hector Keun, h.keun@ic.ac.uk

Metabonomics is a holistic approach to characterizing the time-dependent effects on metabolism of xenobiotic, patho-physiological, or genetic-interventional stimuli. It predominately utilizes NMR spectroscopy of biofluids and tissues, coupled with pattern recognition to automate the interpretation and classification of the observed effects. The approach is increasingly recognized as important by the pharmaceutical industry, especially for predicting the likely toxicity of candidate drugs. The Consortium for Metabonomic Toxicology (COMET) aims to construct databases of xenobiotic toxicity using ca. 100,000 ^{1}H-NMR spectra of biofluids of laboratory rats and mice treated with model compounds, and thus to develop an expert system predictive of the site and mechanism of toxic effects. Here we present data from COMET, applying Soft Independent Modeling of Class Analogy (SIMCA) to toxicity classification, and describe a novel, geometry-based, modeling strategy, Scaled-to-Maximum Aligned Reduced Trajectories (SMART), designed to improve the robustness and predictive success of metabonomic models. SMART was conceived from the hypothesis that the geometry of the time-related response to dose in multivariate space (the metabolic trajectory) would reliably discriminate between different toxins.

Urine spectra (630) of rats from 20 different treatment groups were data-reduced by integrating the signal intensity in 0.04ppm wide regions to produce 205 spectral descriptors. All but one group showed distinct spectral differences from normal physiological variation in rat urine. The resulting reduced spectra were normalized to account for gross concentration differences and regions corresponding to non-endogenous metabolites (~4% of the data) were treated as missing values. 20 PC models were built from ~50% of the data from each of the groups, with the remaining 50% used as a prediction set. A SIMCA methodology was applied, where samples were either assigned to the model that produced the lowest residual, *i.e.* the nearest neighbor according to the distance-to-model, or as unknown if all models give residuals above a critical value. Using the critical distance as an adjustable parameter, ~3/4 of the prediction set could be classified with ~80% accuracy. The integration of SMART-SIMCA produced a significant improvement, classifying 84% of the prediction set with ~90% accuracy, *i.e.* allowed for both higher sensitivity and specificity. Furthermore, SMART-SIMCA required no outlier exclusion during model generation but still retained the facility to detect anomalous samples, demonstrating the robustness of the system.

This work demonstrates that highly predictive models for toxicity classification can be derived from metabonomic data. The SMART strategy has potential benefits for chemometric modeling of any phenomenon that is characterized by highly correlated, multidimensional geometry.

Title: Improved analysis of multivariate data by variable stability (VAST) scaling: Application to NMR spectroscopic metabolic profiling

Authors: Hector Keun, Henrik Antti, Tim Ebbels, Mary Bollard, Olaf Beckonert, Elaine Holmes, John Lindon, Jeremy Nicholson
Biological Chemistry, Biomedical Sciences Division; Imperial College of Science, Technology & Medicine; Sir Alexander Fleming Building, Exhibition Road; South Kensington, London SW7 2AZ; United Kingdom

Keywords: data filtering, variable scaling, metabonomics, VAST, biofluid, NMR, OSC

Presenter: Hector Keun, h.keun@ic.ac.uk

Normal physiological variation in biofluid composition is a serious confounding factor in metabolic profiling research. In addition, any analytical method will introduce artifactual variation that must be separated from the differences of interest. The Consortium for Metabonomic Toxicology (COMET) aims to construct databases of xenobiotic toxicity using ca. 100,000 ^{1}H-NMR spectra of biofluids of laboratory rats and mice treated with model compounds, and thus to develop an expert system predictive of the site and mechanism of toxic effects.

The ability to detect reliably treatment-related responses over significant background noise is important to our metabonomic analyses as for many other analytical applications. Here we present a new pre-processing method, variable stability (VAST) scaling that seeks to minimize the non-systematic variation within classes, and consequently improves the discrimination between clusters, interpretation of variable importance and model predictivity. Applications to data mining for urinary biomarkers of a surgical procedure (unilateral nephrectomy) and to combining multi-site plasma studies of hepatotoxicty are demonstrated. In the former, VAST scaling improved the distinction between operated and non-operated groups and more clearly identified the responsible factors, while in the latter it reduced the influence of a site-dependent contaminant and other baseline inconsistencies.

The beneficial effects of VAST scaling were compared to Orthogonal Signal Correction (OSC), another pre-processing procedure based on data filtering. Both approaches led to similar enhancements for the datasets tested. VAST scaling can be applied to any multivariate dataset that can be classified using prior knowledge.

Title: Improvement of sliding-window gene-shaving clustering for gene expression data

Authors: Young-Hyun Ko, Hyeseon Lee, Chi-Hyuck Jun
POSTECH; Department of Industrial Engineering; Pohang University of Science and Technology; San 31 Hyoja-dong; Pohang 790-784; Republic of Korea

Keywords: sliding-window gene-shaving, gene expression

Presenter: Young-Hyun Ko, notime@postech.ac.kr

Gene-shaving (Hastie *et al*, 2000) is a useful method to extract a meaningful group of genes when the variation of expression is large across samples or conditions. Applied gene-shaving for whole gene data once, we may not find another low-expressed but coherent genes. Choi et al. (2001) propose the sliding-window gene-shaving which is to apply gene-shaving in each window sliding by a certain step size. The final core genes are gathered in the way that taking the cluster with larger gap statistics within a sliding, and taking the union of sets with core genes between sliding. The performance of sliding-window gene shaving depends on the choice of some parameters such as minimum gap statistics, window size and step size.

In this work, we examine how well defined the coherent gene groups by the parameter sets and suggest the general rule for the choice of parameters. Also we modify the step of sorting genes by similarity before applying gene-shaving for saving computational load. With parameters set by the proposed rule and modifying sorting step, the sliding-window gene shaving is improved in terms of accuracy and computational time finding the meaningful groups from gene expression data.

To evaluate the proposed rule and technique, we use several artificial data based on a published data, budding yeast (Chu et al, 1988).

1) T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, P.O. Brown. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns.Genome Biology 2000;2(2).
2) D. Choi, H. Lee, C.-H. Jun. On Combining Clustering Methods for Micoroarray Data Analysis, The Proceedings of International Statistical Institute, 2001.
3) S. Chu, J. Derisi, M. Eisen, J. Mulhalland, D. Botstein, P.O. Brown, I. Herskowitz. The Transritional Program of Sporulation in Budding Yeast. Science 1998;282.

Title: Analysis and monitoring of batch processes using projection methods: An evaluation of alternative approaches

Authors: Yaqiu Chen, Theodora Kourti, John MacGregor
McMaster Advanced Control Consortium; Department of Chemical Engineering;
McMaster University; Hamilton, Ontario L8S 4L7; Canada

Keywords: batch process, historical data analysis, process monitoring, multiway PCA (MPCA), multiway PLS

Presenter: Theodora Kourti, kourtit@mcmaster.ca

Several approaches based on Multi-way PCA and PLS have been proposed for modeling data from batch processes. The models (depending on the way they were derived) are used to analyze historical data for troubleshooting or for process monitoring and fault detection and identification. The main differences of these approaches are the way they unfold the three way data matrix and whether they mean-center the data. The objective of this work is to take a critical look at these approaches, analyze their performance and discuss their advantages and disadvantages.

The approach proposed by Nomikos & MacGregor (MPCA) [1-3] unfold the data matrices batch-wise and then compare batches by analyzing the dominant variation of all variables about their mean trajectories and over the entire history of the batch. Some recent approaches [4] unfold the data matrices variable-wise and use PCA / PLS to extract the dominant variation among the variables at each instant of time without subtracting the mean trajectories (WPCA). For sake of completeness an intermediate approach (Alternative Multi-way PCA, AMPCA) is introduced, in which the data matrices are unfolded variable-wise, but only after the mean trajectories have been removed. The dominant variation in the deviation of the variables from their average trajectories at each instant of time is then extracted via PCA.

Examining these three formations allows one to easily contrast the impact of the different steps used in the multi-way approaches for the analysis of historical batch data and the monitoring of new batches. The two key steps:

- ¥ the removal of the mean trajectories and
- ¥ the nature of the unfolding

are shown to have a major impact on the subsequent use of the PCA models. In particular, deviations among batches can only be clearly revealed if the unfolding is done batch-wise and if the mean trajectories are removed. Furthermore by unfolding variable-wise, only the covariance structure among the variables at each instant of time is extracted and the time history of this variation is ignored. As a result, the batch data have to be broken into many phases with separate models considered for each phase. On the other hand, for process monitoring the variable-wise unfolding does not require a procedure for reconstructing the missing data over the remainder of the batch.

These aspects are discussed from a basic modeling point of view and are contrasted using several industrial batch data sets.

1) P Nomikos, JF MacGregor. Monitoring Batch Process Using Multiway Principal Component Analysis. AIChE J 1994;40:1361-1375.

2) P Nomikos, JF MacGregor. Multi-way Partial Least Squares in Monitoring Batch Process. Chemom Intell Lab Syst 1995;30:97-108.

3) T Kourti, JF MacGregor. Tutorial: Process analysis, monitoring, and diagnosis, using multivariate projection methods. Chemom Intell Lab Syst 1995;28:3-21.

4) S Wold, N Kettaneh, H Frid n, A Holmberg. Modeling and diagnostics of batch process and analogous kinetic experiments. Chemom Intell Lab Syst 1998;44:331-340.

Title: Data processing strategies for the determination of glucose from low resolution Fourier-transform near-IR spectra

Authors: Kirsten Kramer, Gary W. Small
Department of Chemistry & Biochemistry; Clippinger Labs; Ohio University;
Athens, OH 45701-2979; USA

Keywords: FT-IR, glucose, resolution, multivariate calibration

Presenter: Kirsten Kramer, kk398093@ohio.edu

Current at-home methods for measuring blood glucose levels are based on test-strip technology that requires the collection of a blood sample. This process can be inconvenient and painful for diabetics who must monitor themselves five to six times daily. A strong desire to develop noninvasive blood glucose monitoring capabilities has led to preliminary methodologies focused largely on vibrational spectroscopy, using either transmission or reflectance measurements. These approaches require multivariate calibration techniques to aid in the extraction of small glucose signals from the overlapping background absorbance arising from the complex sample matrix.

In our laboratory, Fourier transform infrared (FT-IR) transmission measurements of the near-infrared combination band region (4000 — 5000 cm^{-1}) are being evaluated for possible use in a noninvasive glucose analysis. The principal challenges to this approach are:

1) the fact that a clean (*i.e.*, glucose-free) spectral background will not be measurable,
2) the glucose signal itself is several orders of magnitude smaller than the overwhelming background absorbance of water,
3) other matrix constituents (*e.g.*, proteins, triglycerides, etc.) also have significant background absorbance, and
4) current FT-IR instrumentation is relatively fragile and not affordable to the typical home user.

The first three issues can be addressed with appropriate data processing methods such as signal processing and multivariate calibration. However, the final issue must be addressed through building a more rugged, less expensive instrument in which spectral resolution is compromised (*i.e.*, a shorter mirror movement in the interferometer is employed).

This presentation will focus on the impact of spectral resolution on the glucose analysis and the development of appropriate data processing protocols to allow well-performing calibration models to be developed with low-resolution spectra. Issues such as the utility of preprocessing tools and the choice of calibration methodology will be evaluated. This work will be performed with FT-IR spectra of samples of glucose in a synthetic biological matrix consisting of varying levels of triacetin (simulant for triglycerides) and bovine serum albumin (simulant for human blood proteins) in an aqueous buffer.

Title: Genetic algorithms for pattern recognition and multivariate calibration: I. A solution to the class membership problem

Authors: Barry K. Lavine, Charles E. Davidson
Department of Chemistry; Clarkson University; Potsdam, NY 13699-5810; USA

Keywords: GA, data mining, PCA, nonlinear pattern recognition, Kohonen NN, machine learning

Presenter: Barry Lavine, bklab@clarkson.edu

We have developed and tested a genetic algorithm (GA) for pattern recognition, that identifies features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. The principal component analysis routine embedded in the fitness function of the pattern recognition GA acts as an information filter, significantly reducing the size of the search space since it restricts the search to features whose principal component plots showed clustering on the basis of class. In addition, the algorithm focuses on those classes and samples that are difficult to classify as it trains using a form of boosting to modify the class and sample weights. Boosting minimizes the problem of convergence to a local optimum because the fitness function of the GA changes as the population evolves towards a solution. Samples that consistently classify correctly are not as heavily weighted as those samples that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The fitness function of the pattern recognition GA has been generalized to tackle messy pattern recognition problems, *e.g.* outliers and/or nonlinear relationships in data, by using a Kohonen neural network in the Toroidal configuration in lieu of principal component analysis. Feature subsets selected by the GA are projected onto a curved surface, allowing nonlinear sources of variation in the data to be modeled. The Kohonen neural network has the added advantage of being insensitive to outliers.

Title: Genetic algorithms for pattern recognition and multivariate calibration: II. Transverse learning

Authors: Barry K. Lavine, Charles E. Davidson
Department of Chemistry; Clarkson University; Potsdam, NY 13699-5810; USA

Keywords: GA, transverse learning, PCA, Hopkins statistic, feature selection

Presenter: Barry Lavine, bklab@clarkson.edu

A genetic algorithm that uses a fitness function, which incorporates transverse learning, has been developed by coupling the Hopkins statistic to the original fitness function of the pattern recognition GA (described in the previous paper). The Hopkins statistic searches for features that increase the clustering of the data where the fitness function of the pattern recognition GA identifies feature subsets that create class separation. Scaling the Hopkins statistic using a sigmoid transfer function and applying an influence function to deweight observations with high leverage and thereby robustify the Hopkins statistic ensures that our modified Hopkins statistic is a meaningful metric to assess clustering. We will be able to explore the structure of a data set, for example, discover new classes, by simply tuning the relative contribution of the Hopkins statistic and the original pattern recognition fitness function to the overall fitness score. For training sets with small amounts of labeled data and large amounts of unlabeled data, this approach is preferable as our previous have shown since the information in the unlabeled data is used by the fitness function to guide feature selection and prevent over-fitting. Using this approach, feature subsets will be selected to optimize clustering and to maximize the distance between the different classes in the data set; thereby ensuring that our GA will perform better than a learning model developed from a set of features whose selection is based solely on the dichotomization power of features for the labeled data points. This approach to feature selection is an example of semi-supervised learning because it incorporates aspects of supervised and unsupervised learning to develop a new paradigm for multivariate data analysis where classification, clustering, feature selection, and prediction are combined into a single step enabling a more careful analysis of data. The semi-supervised learning approach has been extended as part of our research on using chemometrics for discovery to include problems in multivariate calibration.

Title: Models for measurement error covariance in multichannel instrumentation

Authors: Marc N. Leger, Peter D. Wentzell
Department of Chemistry; Dalhousie University; Halifax, Nova Scotia B3H 4J3; Canada

Keywords: multivariate calibration, error covariance, preprocessing

Presenter: Marc Leger, mleger@chem1.chem.dal.ca

Although measurement uncertainty is an integral part of chemical measurements, most multivariate calibration methods ignore this important factor by implicitly assuming uniform, uncorrelated noise or attempting to accommodate the noise structure through routine preprocessing. Inaccurate assumptions about measurement error covariance structure can lead to suboptimal multivariate calibration models. Maximum likelihood principal components regression (MLPCR) is a multivariate calibration method that takes into account the measurement error structure of analytical data, and has been shown to provide lower prediction errors when data exhibit heteroscedastic and correlated error.

A drawback of MLPCR is that the measurement error covariance matrix has to be estimated for every new application, or in some cases, for every sample. Sample replicates are necessary to generate this error covariance matrix, but it can be inconvenient or even impossible to obtain these. However, it has been observed that some multichannel instruments present characteristic error covariance structures, regardless of the sample being analyzed. In these cases, it may be possible to describe the error covariance matrix using a minimal number of factors related to the particular instrument or sample.

The objective of this work is to model measurement error covariance for typical multichannel instrumental methods. By understanding the behavior of error covariance for spectroscopic instruments such as near-IR, UV/VIS and fluorescence instruments, it may be possible to incorporate covariance structures into a pre-processing step or MLPCR, thus greatly reducing the need for measurement replicates. Widely used pre-processing methods such as scaling, filtering and multiplicative signal correction (MSC) are often associated with specific instruments, and a better understanding of measurement error covariance matrices may enhance our understanding of this relationship. Covariance models will be established based on experimental data of different types, and the implications of these to multivariate calibration will be discussed. Furthermore, error covariance matrices estimated from a few samples may have substantial errors that can offset the advantages of using the information in the first place.

Title: Development applications of guided microwave spectroscopy using chemometrics

Authors: Victoria C. Loades, Anthony D. Walmsley
Department of Chemistry; University of Hull; Cottingham Road;
Kingston-upon-Hull HU6 7RX; United Kingdom

Keywords: process analysis, spectroscopy, microwave

Presenter: Victoria Loades, v.c.loades@chem.hull.ac.uk

As process analysis has begun to mature, with the use of analyzers now being quite common place in many industries, there is sufficient interest to now turn to more novel alternative methods, such as Raman spectroscopy, laser induced fluorescence (LIF), cavity ring down spectroscopy and microwave spectroscopy. Process analyzers need to be able to monitor the entire spectrum of applications, rather than simple gas or liquid phases, however very few techniques can deal with multiphase solutions. The benefit of microwave spectroscopy is that analyses the entire sample (it does not use a probe) and is suitable for the analysis of non-homogeneous substances, it is non-invasive and non-destructive.

A microwave spectra comprise of two fundamental properties: a dielectric loss due to a reduction in wave velocity as it passes through a given sample and a dielectric constant from a reduction in magnitude as a result of loss of energy due to friction as molecules orientate in the field. These properties are sample dependent and typically free from interferants such as particle size and interfaces. The resulting spectra are broadband and overlapping. Unlike traditional spectroscopies where the peak of interest will vary in height as concentration increases, for microwaves the entire spectra will shift and often change shape significantly.

This paper investigates the application of chemometric modeling and demonstrates that it is feasible to use process microwave analyzer as an alternative to other more established methods and that in some cases demonstrating a significant benefit. We will present some applications of guided microwave spectroscopy (GMS) combined with chemometric modeling that demonstrate the versatility of this technique for process monitoring, examples including the measurement of acetonitrile and ethanol in water [1], binary alcohol mixtures and the analysis of moisture in tobacco [2] and a recent real life application, the monitoring of multiphase (solid, organic and aqueous) industrial process samples, that is a true test of the analyzers potential.

1) A.D. Walmsley and V.C. Loades, Analyst 2001;4:417-420.

2) A.D. Dane, G.J. Rea, A.D. Walmsley, SJ. Haswell. Anal Chim Acta 2001;429:187-196.

Title: Methods for luminescence lifetime determination

Authors: Christina M. McGraw, Gamal Khalil, James B. Callis
Department of Chemistry; University of Washington; Box 351700;
Seattle, WA 98195; USA

Keywords: luminescence, lifetime, phosphorescence, fluorescence

Presenter: Christina McGraw, cmcgraw@u.washington.edu

Despite the usefulness of luminescence lifetimes as an analytical technique there seems to be no consensus on the best method for lifetime determination. Workers in this field split into two camps: those who favor methods that proceed in the time domain and those who favor methods that proceed in the frequency domain.

We have carried out a careful comparison of these methods using the same samples, exciting them with the same light source, and using the same detector. In this study we subject the system to various optical excitation sequences in order to identify the optimal method for measuring the luminescence lifetime. The study is limited to the case of stimulation with LEDs or solid state lasers such that no trade off is possible between duty factor and instantaneous intensity. Moreover, we are not limited to impulse and sine wave excitation, but can evaluate chirp and random excitation sequences as well.

Title: Missing data estimation based on a combined wavelet-PCA technique

Authors: Vitor V. Lopes, Jose C. Menezes
Center for Biological & Chemical Engineering; Technical University of Lisbon;
Av. Rovisco Pais; P1049-001 Lisbon; Portugal

Keywords: data reconstruction, wavelets, PCA, industrial processes

Presenter: Jose C. Menezes, bsel@ist.utl.pt

Industrial data sets most often contain missing data. This might be due to measurement instrument failure or sensor saturation. In this paper a non-iterative technique will be presented in order to estimate missing data values. This technique can be applied to historical data in which all variables fail (*e.g.*, when data-logging fails to communicate with the field bus).

The proposed technique assumes that some serial correlation exists between data records and variables and is based on the combined application of wavelets and principal component analysis to the incomplete datasets. The simultaneous application of a wavelet projection matrix (W) and the PCA projection matrix (L) in order to compute a score matrix (T) yields $T_{scores} = W\,?X\,?L$. By using the vectorization operator it is possible to reconstruct the original data set as: $vec(X) = (L \quad W^{-1})vec(T_{scores})$. The crucial part of the process is the estimation of the score matrix (T) because the estimation of the missing data depends on it.

The wavelet projection matrix (W) only depends on the wavelets selected and on the selected decomposition level. Its elements are constructed independently of the dataset. The PCA projection matrix (L) depends on the covariance matrix that is constructed by computing each covariance pair independently (by doing this maximum use of available the data is made since the number of missing points will be minimal in the covariance calculation for each pair of variables).

The scores matrix (T) is then estimated by solving the linear system using only the know data points: $vec(X)_{good} = (L \quad W^{-1})_{good}\, vec(T_{scores})$. The estimation of the missing data can be done afterwards by using the previously estimated scores (T): $vec(X)_{bad} = (L \quad W^{-1})_{bad}\, vec(T_{scores})$. The computation process is linear and it doesn t need any iteration step. The type of wavelets selected and the selected decomposition level must be chosen in a way that the condition number for the matrix $(L \quad W^{-1})_{good}$ is the lowest possible. It was found that the biorthogonal Villasenor wavelets ([3 7] type) were a good choice and the decomposition level was dependent on the maximum length of the consecutive missing data.

The paper describes the application of the proposed technique to datasets from an industrial petrochemical unit, showing excellent reconstruction capabilities of the missing data.

Title: Determination of sulphate in extracted phosphoric acid using flow injection enthalpimetric analysis

Authors: Q.Y. Zhang, Mengqiang Wu
Institute of Microelectronics and Solid State Electronics;
University of Electronic Science and Technology of China; P.O.Box 1503-2;
Chengdu 610054; People's Republic of China

Keywords: sulphate determination, extracted phosphoric acid, flow enthalpimeter, flow injection analysis

Presenter: Wu Mengqiang , mwu@uestc.edu.cn

In the manufacture of extracted phosphoric acid from the acidolysis of phosphoric ores by sulphuric acid, the measurement of sulphate or sulphur trioxide of the products is an essential control analysis that is required at various times because of the importance of sulphur trioxide upon the quality and the expense for the products. It s therefore desirable to have a simple and fast approach of analysis that can be used on line by the plant-operating personnel.

The optimum conditions for the adaptation of the enthalpimetric method for sulphate in extracted phosphoric acid in plants to a flow injection system are described. The concentration of sulphate can be determined in sulphur trioxide with the detection limit of 3.2 g/L, the relative standard deviation of 1.6% and the recovery of 96~104% at a rate of about 75 samples per hour. The co-existing substances of possible content in extracted phosphoric acid cause no significant interference but phosphoric acid, the interference of which is capable of being suppressed easily.

The results of the analysis of samples from various plants by the proposed procedure can be regarded as satisfactory in comparison with those obtained by the volumetric method. The presented method is suitable for routine determination of sulphate in extracted phosphoric acid in plants. The precision and accuracy of the enthalpimetric method are lower than those of the classical volumetric methods, whereas main advantage (essential for control analysis in plants) of the proposed method are that the inexpensive flow enthalpimeter required, as well as the rest of the apparatus, can be maintained easily. It is suitable for on-line automatic analysis without additional samples preparatory steps.

Title: Fuzzifying classification trees for improved prediction

Authors: Anthony Myles, Steven D. Brown
Department of Chemistry & Biochemistry; University of Delaware;
Newark, DE 19716; USA

Keywords: classification trees, feature selection, missing data

Presenter: Anthony Myles, myles@udel.edu

Classification trees offer many advantages over other classification techniques including inherent feature selection, missing value handling, incorporation of prior knowledge, and their ability to handle a variety of data types. Also, their flexible and hierarchical structure allows classification trees to tackle complex problems, while still maintaining simple interpretation. However, some problems exist. Regions surrounding decision boundaries within a classification tree are often unstable and easily shifted by the addition of a few samples. Therefore the classification of samples near a decision boundary may change with only small differences in sample measurements. Closeness to a decision boundary affects the degree of confidence one has that a sample should follow a specific path.

To relieve some of the decision boundary instability, crisp splits can be replaced by fuzzy splits. A fuzzy split can best be explained as a splitting region, described by some function, where samples have partial membership to each partition. A new method for creating fuzzy splits is described that specifically relates the region stability to partition membership. The fuzzy split is created by randomly selecting a subset of samples present in the local data space, building a distribution from the optimal split values for each iteration, and then estimating the parameters for the fuzzy split. This new method creates a classification tree that is more generalized to the data, provides information about variable importance and selection, improves surrogate splitting (missing value handling), and supplies probabilistic class prediction. This method will be described in detail and applied to standard data sets.

Title: Elucidation of the structure of a protein folding intermediate (molten globule state) using multivariate curve resolution alternating least squares

Authors: Susana Navea, Anna de Juan, Rom Tauler
Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

Keywords: protein folding, MCR-ALS, three-way, modelling of reaction intermediates, curve fitting, circular dichroism, IR

Presenter: Susana Navea, susana@apolo.qui.ub.es

It is well known that proteins are not folding at random and different mechanisms have been proposed to describe this process. Proteins can be denatured by changes on the temperature, pH or concentration of denaturant agents. Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) has been used to study the protein folding process.

This resolution method provides a complete description of the process, much more informative than the usual biochemical approaches, based either on the deconvolution of individual spectra or on one-wavelength process monitoring. The pure resolved concentration profiles contain information about the mechanism of the process and the evolution of each of the protein conformation involved. The spectral shapes of the pure spectra relate to the abundance of protein structural motifs (helices, sheets,), thus characterizing the protein conformation involved.

Thermally induced changes in the secondary and tertiary structure of a-apolactalbumin have been monitored using far- and near-UV circular dichroism (CD) and infrared (IR) spectroscopies. The analysis of the near-UV CD measurements, used to monitor changes on the tertiary structure of proteins, detected only two species, related to the native and unfolded states of this protein structural level. The analysis of the far-UV CD and IR measurements, used to monitor changes on the secondary structure of proteins, detected the same number of species linked now to the native and unfolded states of the secondary structural level. Only the simultaneous analysis of all data matrices, that it is to say, the analysis of a row-wise augmented matrix formed by different spectroscopic measurements, monitoring changes on the tertiary and secondary structure allowed for the detection of the three protein conformations involved in the folding of a-apolactalbumin: native, unfolded and molten globule state because they present a different row-wise augmented spectra. The molten globule state is an intermediate that presents an unordered tertiary structure and a native-like secondary structure, which can be confirmed from the shape of the resolved spectrum for this species. Molten globule states cannot be isolated experimentally because they are thermodynamically unstable; therefore, MCR-ALS has allowed to resolve the structure of an intermediate conformation that otherwise would have been difficultly detected and, by no means, characterized. The effect of ionic strength on the occurrence and evolution of the intermediate species has also been assessed.

To complete the study, the assignment of the percentage of the different secondary structure motifs to each protein conformation detected is carried out. To do so, the pure resolved far-UV CD spectra are modeled with the help of a reference data set, which links protein CD spectra to their elucidated X-ray crystallographic secondary structures, using least squares curve fitting approaches.

Title: Classification with boosted multilayer perceptrons

Authors: Matteo Pardo, G. Sberveglieri, G. Valentini, F. Masulli
INFM & University of Brescia; Via Valotti, 9; I25133 Brescia; Italy

Keywords: classification, resampling methods, NNs, electronic nose

Presenter: Matteo Pardo, pardo@tflab.ing.unibs.it

The use of a single neural network (normally a multilayer perceptron) as a classifier is a common solution to pattern recognition problems in many application fields, comprising analytical chemistry. A direction in which research in supervised learning is making great progresses is the study of techniques for combining the predictions of multiple classifiers (briefly called ensembles) to produce a single classifier. The resulting classifier is generally more accurate than any of the individual classifiers making up the ensemble. Both theoretical and empirical research has demonstrated that a good ensemble is one where the individual classifiers are both accurate and make errors on different parts of the input space. Two popular methods for creating accurate ensembles that emerge from the recent machine learning literature are Bagging and Boosting. These methods rely on resampling techniques to obtain different training sets for each of the classifiers.

In this paper we apply Boosting to the classification of data collected with the Pico-1 Electronic Nose (EN) developed at the Gas Sensor Lab in Brescia. ENs, in the broadest meaning, are instruments that analyze gaseous mixtures for discriminating between different (but similar) mixtures and, in the case of simple mixtures, quantify the concentration of the constituents. ENs basically consist of a sampling system, an array of chemical sensors, electronic circuitry and data analysis software.

Experiments were performed on two groups of coffees, consisting respectively of 7 different blends (containing the Italian Certified Espresso (ICE)) and of 6 single varieties (SV) plus the ICE. The results obtained with boosting are compared with those obtained (without boosting) with the cascade of PCA and multilayer perceptrons. The boosting algorithm was able to halve the classification error for the blends data and to diminish it from 21% to 18% for the more difficult monovarieties data set with confront to a single learner.

It will be also shown that the test error on the two data sets continues to decrease, even after the training error reaches zero. This fact has been already observed in the literature on boosting and has been explained in the framework of large margin classifiers, interpreting boosting as an algorithm that enlarges the margins of the training examples. Even if the training error is near to zero, the boosting algorithm continues to enhance the margins, focusing on the hardest examples. As a consequence, the generalization capabilities of the boosted ensemble are improved.

Title: Multiway data analysis of environmental contamination sources in surface natural waters of Catalonia, Spain

Authors: Emma Per -Trepat [1], M nica Flo[2], Montserrat Munoz [2], Elisabeth Teixid [2], Merce Figueras[2], Lourdes Olivella[2], Manel Vilanova[2], Josep Caixach[3], Antoni Ginebreda[2], Rom Tauler[1]

1 Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

2 Agencia Catalana de l'Aigua; Provenca 204-208; E08036 Barcelona; Spain

3 Department of Ecotechnologies; IIQAB-CSIC; Jordi Girona 18-26; E08034 Barcelona; Spain

Keywords: PCA, MCR, environmental contamination, surface waters, Catalonia

Presenter: Emma Pere-Trepat, emma@apolo.qui.ub.es

Analytical data acquired in current environmental natural water monitoring programs are usually compiled in spreadsheet-like large data tables. These tables include the determination of analytical concentrations of selected organic and inorganic chemical compounds measured on multiple samples including natural water samples, sediment samples and/or different type of organisms (*i.e.*, fishes) samples. Different sampling sites at different geographical locations and different sampling campaigns and/or sampling time periods are considered. The whole data set may be arranged according to these data features. The simplest possible data arrangement is a single data table or data matrix giving the concentrations of one particular set of chemical compounds (organic or/and inorganic) measured on different sampling sites of the same type (water, sediment or organism/fish) and at the same sampling campaign or sampling period. Bilinear data decomposition of this data matrix using either Principal Component Analysis or Multivariate Curve Resolution will give score and loading matrices describing the composition profiles of the main contamination sources and their geographical distribution. More complex data arrays are possible including the concentration variations over different types of samples (water, sediments and organisms) and of different sampling periods and campaigns. These more complex data arrays may be arranged in column- and row-wise augmented and superaugmented data matrices and in data cube, or hypercubes or higher order data tensors. Bilinear decomposition of these more complex data structures will provide additional information about main temporal distribution and accumulation of contamination sources in the different type of samples (water, sediments, organisms/fishes).

In the frame of an extensive multi annual environmental monitoring program from the Catalan Water Agency (Ag ncia Catalana de l'Aigua), a very large number of samples from the whole geographical area of Catalonia (Spain) have been analyzed. River and lake water samples, fish samples (like barb , bagra comuna , bleak, carp and trout) and sediment samples were included in this study. Organic microcontaminants like organochlorine compounds (such as HCB, DDT, DDE, DDD, PCBs, etc.), commonly found polycyclic aromatic hydrocarbons (PAHs) and main inorganic contaminants like heavy metal ions were included. Main contamination sources and their geographical and temporal distributions as well as their soil and biological accumulations are estimated. Global environmental quality of surface natural waters in the whole Catalonia area is critically evaluated.

Title: Identification and distribution of microcontaminant sources of nonionic surfactants, their degradation products, and linear alkylbenzene sulfonates in coastal waters and sediments in Spain by means of chemometric methods

Authors: Emma Per -Trepat [1], Mira Petrovic[2], Dami Barcel [2], Rom Tauler[1]

1 Chemometrics Group; Departament de Qu mica Anal tica; Universitat de Barcelona; Diagonal, 647; E08028 Barcelona; Spain

2 Department of Environmental Chemistry; IIQAB-CSIC; Jordi Girona 18-26; E08034 Barcelona; Spain

Keywords: PCA, Multivariate Curve Resolution Alternating Least Squares (MCR-ALS), surfactants, environmental contamination

Presenter: Emma Pere-Trepat, emma@apolo.qui.ub.es

Principal Component Analysis (PCA) and Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) methods are proposed for chemometric analysis of large multivariate data tables obtained in environmental studies. In this work, both methods are extended and compared for the analysis and interpretation of data sets obtained in environmental programs, covering a large number of chemicals analyzed regularly and simultaneously in many samples and in widespread geographical areas. Eighteen chemical compounds including nonionic surfactants, their degradation products, and linear alkylbenzene sulfonates were included. A total number of 74 samples, (35 water samples and 39 sediments samples), corresponding to different months, from March 1999 to July 2000, and in 32 different geographical sites located along the Mediterranean Spanish coast (such as Barcelona, Tarragona, Almeria harbor and Bay of Cadiz) were analyzed for these eighteen chemical compounds. Main contamination environment sources of these compounds were identified and interpreted according to their chemical composition and according to their resolved geographical and temporal distribution profiles.

Title: Determination of doxorubicin in human plasma by excitation-emission matrix fluorescence and multi-way analysis

Authors: Marcello Trevisan, Ronei J. Poppi
Instituto de Qu mica - UNICAMP; Universidade Estadual de Campinas;
Campinas, S o Paulo 13081-970; Brazil

Keywords: doxorubicin, EEM fluorescence, multi-way

Presenter: Ronei Poppi, ronei@iqm.unicamp.br

Doxorubicin (DXR) is a cytotoxic anthracycline antibiotic widely employed in clinical practice and has significant antitumor activity against several human malignancies including leukemia and breast cancer. Quantitative determination of this compound, mainly in plasma patients, is required since toxic effects are commonly observed after prolonged treatment. Methods based on chromatographic separations are normally employed, but they are laborious and time consuming. DXR is an amphiphilic molecule and it has a fluorescent hydroxy-subtituted anthraquinone chromophore. Direct fluorescence drug measure in plasma can be an alternative method to chromatographic separation due its high sensitivity and selectivity allied to fast and clean spectra acquisition. However, utilization of a single emission wavelength limits the quantitation due the native plasma fluorescence. A procedure to overcome this drawback is the utilization of the excitation-emission matrix fluorescence of several samples and chemometric multi-way methods.

Twenty-two human plasma samples were obtained from healthy volunteers and solutions were prepared by dissolving the appropriate quantity of Doxorubicin hydrochloride in 2mL of plasma. DXR solutions in the range of 0.75 to 15.00 ng/mL were prepared. The EEM spectra were recorded at excitation wavelengths from 390 to 500 nm at regular steps of 2 nm and the emission wavelengths from 510 to 650 nm at 0.5 nm.

PARAFAC (parallel factor analysis) was employed to EEM spectra deconvolution and possible DXR quantitation. From PARAFAC calculations it was observed that two factors described more than 99.9% of data variance with a core consistence of 99.8%. From the emission spectra profiles obtained in the decomposition, it was possible to see that one spectrum was due the DXR and another was the plasma. Using the loadings of samples mode, a calibration was performed against the DXR concentrations, resulting in a RMSEP of 0.53 ng/mL.

N-PLS was employed to DXR quantitation and cross-validation was used to find the optimum number of factors. The optimum number of factors was three, resulting in a RMSEP of 0.30 ng/mL. From the results obtained it was possible to conclude that the proposed methodology can be used to DXR quantitation in human plasma.

Title: Handling large datasets in the food industry

Authors: Vibeke T. Povlsen, Frans W.J. van den Berg
Department of Dairy and Food Science; Royal Veterinary and Agricultural University;
Rolighedsvej 30, iii; DK1958 Frederiksberg C; Denmark

Keywords: large datsets, multiblock, uncertainty, jack-knife, prediction error

Presenter: Vibeke T. Povlsen, vip@kvl.dk

In the food industry large amounts of data from different sources (analytical chemical measurements, rheological intermediate and product characterization, production or storage process settings, etc.) are collected to study, optimize and control frequently complicated food production processes. Data handling and analysis of these large data-sets can be a challenging job. The primary question in explorative process monitoring is to scan the constituting data-blocks, organized in some meaningful structure, separating the informative from the uninformative contributions. One way of handling this problem is a so-called multiblock approach. The multiblock factor model provides information about the role of the individual blocks (*e.g.*, one type of measurement) in the overall model, given by the super/block weight of the model, where the super weight shows the influence of each block for each factor.

The multiblock approach might be considered as a preprocessing step since the non-informative blocks can be left out based on the information given by the super weight. The difficulties can be to select the appropriate multiblock model. Several issues have to be taken into consideration when determining the right model complexity and block structure. The weighting of each block influences the complexity and the stability of the final model. Usually the decision is based on the prediction error, the explained variance (of predictor and response blocks) and predictive ability. Normally the aim is to choose the model with the lowest prediction errors. But in some cases this might lead to over-fitting, non-parsimonious models, whereas the more appropriate complexity would be around the first local minimum of the prediction error, when the changes are no more significant. The problem is even more pronounced for the often limited number of object/samples (*e.g.*, pilot plant experimental runs) available in the exploratory stage for product or process development.

In the present work, an illustration of the determination of the correct model complexity using a jackknife resampling based approach for the uncertainty of the prediction error and block weighting is given. The prediction error and the uncertainty interval of this prediction error indicate the regions where the model gives a prediction error that is satisfactory low and accurate in relation to the overall model stability. Emphasis is placed on the graphical diagnostics, significantly simplifying the model interpretation, thereby assisting food engineers in their tasks.

Title: Unmixing complex chromatograms

Authors: Scott Ramos, Marlana Blackburn, Brian Rohrback
Infometrix, Inc.; PO Box 1528; Woodinville, WA 98072; USA

Keywords: mixture analysis, chromatography, petroleum, MCR-ALS

Presenter: Scott Ramos, scott_ramos@infometrix.com

The goal of this presentation is to locate the end member profiles given a set of mixtures of multi-peak chromatograms. The chemometrics literature identifies several procedures to accomplish the goals of mixture analysis: from a set of bilinear data, extract estimates of the underlying pure profiles in the two dimensions. Traditionally, mixture analysis has been employed in hyphenated techniques to characterize unresolved constituents. For example, in a GC/MS experiment, when unresolved chromatographic peaks occur, the goal is to discover the elution profiles of the actual components, along with the corresponding mass spectra.

The same approach can be used to separate pure chromatographic profiles from a series of mixed source chromatograms. The bilinear matrices of this work are composed of a composition dimension — the chromatographic profiles — and a concentration dimension — derived from the relative contributions of the end members. Methods for mixture analysis have advantages and disadvantages; two popular methods were chosen for evaluation: multivariate curve resolution and alternating least squares. Examples drawn from the petroleum world will illustrate the methods.

Title: Theoretical and chemometric interpretation of optical low-coherence reflectometry experimental studies of polystyrene nanospheres

Authors: Summer Lockerbie Randall, Anatol M. Brodsky, Lloyd W. Burgess
CPAC; University of Washington; Box 351700; Seattle 98195-1700; USA

Keywords: optical low-coherence reflectometry, highly scattering, PCA

Presenter: Summer Lockerbie Randall, sr33@u.washington.edu

Optical Low-Coherence Reflectometry (OLCR) is a white-light interferometric technique used to evaluate particle size and concentration in highly scattering systems, as well as to monitor the optical thickness of transparent films and coatings. However, models to extract sample information from the decay profiles of highly scattering samples have not been satisfactorily developed.

Recent work has focused upon the deconvolution of sample parameters from the scattering system profiles of standard polystyrene nanospheres through further development of multiple light scattering theory. Principal Components Analysis will be used as an independent tool to study quantitative relationships between profiles from different standards, and results will be compared to the parameters specified by the developing theory. Interpretation and combination of both analysis methods will be used to create a model applicable to industrially relevant sample types. Principal Least Squares may be used for its predictive capabilities upon unknown samples.

Title: Study of the influence of selectivity and sensitivity parameters on detection limits in multivariate curve resolution of chromatographic second-order data

Authors: M» Jos Rodr guez, Ricard Boqu , F. Xavier Rius
Departament de Quimica; Universitat Rovira i Virgili de Tarragona;
Pla a Imperial Tarr co, 1; E43005 Tarragona, Catalonia; Spain

Keywords: LOD, MCR, second-order, ITTFA

Presenter: M» Jose Rodr guez Cuesta, cuesta@quimica.urv.es

We make available a procedure to calculate limits of detection, LOD, for analytical methods that generate second order data, and that follows the IUPAC recommendations, *i.e.* both type A and B probabilities of error are taken into account. The strategy consists on using iterative target transformation factor analysis (ITTFA) to obtain the LOD estimator from the estimated pure chromatograms and spectra of each component of a mixture. Once chromatograms of standards are estimated, we can represent the height (or the area) of the peak corresponding to the analyte of interest in each standard versus concentration, and establish a linear relationship between them (quantitative analysis). From the univariate regression line, the limit of detection (LOD) can be calculated satisfactorily using classical univariate LOD estimators.

The aim of this work is also to provide indications to the end user about the quality of the LOD estimator as a function of the characteristics of the problem at hand (sensitivity, chromatographic and spectral selectivity) in a typical analytical determination such as high resolution liquid chromatography with diode array detection, and how and to which extent we can improve detection limits by modifying these experimental variables.

Limits of detection calculated by this procedure will be affected by the quality of the resolved profiles. Therefore simulations have been performed according to a factorial experimental design, modifying the parameters that affect the curve resolution method. The results show that chromatographic resolution has the strongest effect, followed by the sensitivity of the signal, and, with a weaker effect, the spectral selectivity.

The proposed strategy to calculate limits of detection when dealing with second order data, has been validated with simulated data and with real data obtained from the analysis of phenolic mixtures containing 2-chlorophenol.

Title: Automated interpretation of multivariate instrument data using ActiveX

Authors: Brian Rohrback, Marlana Blackburn, Scott Ramos
Infometrix, Inc.; PO Box 1528; Woodinville, WA 98072; USA

Keywords: process monitoring, inferentials, steady state

Presenter: Brian Rohrback, brian_rohrback@infometrix.com

Typically multivariate models are created by individuals well-versed in chemometrics. However, once models are validated (*i.e.*, demonstrated to be effective and robust), non-chemometricians should be able to use them to generate predicted properties and associated outlier diagnostics in an automated fashion. We will present an easy-to-use front end configured for a process monitoring exercise that branches through a decision tree, assessing steady-state and computing inferentials. The application communicates with Pirouette using ActiveX.

Title: Multivariate statistical process control for continuous monitoring of an early warning fire detection system

Authors: Renee D. JiJi, Mark H. Hammond, Susan L. Rose-Pehrsson
Chemistry Division, Code 6116; Naval Research Laboratory; 4555 Overlook Avenue; Washington, DC 20375; USA

Keywords: multivariate SPC, fire detection, sensor array, sensor network

Presenter: Susan Rose-Pehrsson, srose@ccs.nrl.navy.mil

An Early Warning Fire Detection (EWFD) system was developed in order to improve the rate of fire detection and reduce false alarms as a part of the Navy program Damage Control-Automation for Reduced Manning (DC-ARM). The EWFD system was comprised of four sensors that included photoelectric and ionization smoke detectors, as well as carbon monoxide and carbon dioxide sensors. Fourteen sensor arrays were distributed in 10 compartments and passageways throughout the ex-USS SHADWELL, the Advanced Damage Control fire research platform of the Naval Research Laboratory. Data was collected for a series of smoldering and flaming fire sources as well as nuisance sources in different compartments over two decks. The network of sensor responses, location and temporal data may be used to determine source location and fire rate of growth. Monitoring of EWFD systems requires a method that can effectively discriminate between nuisances, actual fires and their byproducts in adjacent compartments.

Title: Probabilistic neural network for early fire detection using multi-criteria sensor arrays

Authors: Susan L. Rose-Pehrsson, Mark H. Hammond, Daniel T. Gottuk,
Jennifer T. Wong, Mark T. Wright
Chemistry Division, Code 6116; Naval Research Laboratory; 4555 Overlook Avenue;
Washington, DC 20375; USA

Keywords: multivariate methods, multi-criteria system, fire detection, NN

Presenter: Susan Rose-Pehrsson, srose@ccs.nrl.navy.mil

The Navy program Damage Control-Automation for Reduced Manning is enhancing automation of ship functions and damage control systems. A key element to this objective is the improvement of current fire detection systems. A multi-criteria approach to early warning fire detection (EWFD) has been developed to provide reliable warning of actual fire conditions in less time with fewer nuisance alarms than can be achieved with commercially-available smoke detection systems. Two standard smoke detectors were used as benchmarks to measure the sensor array performance. The EWFD system was comprised of four sensors that included photoelectric and ionization smoke detectors, as well as carbon monoxide and carbon dioxide sensors. A probabilistic neural network was used to discriminate real fires and nuisance sources. Novel methods were used to reduce the number and to select optimal patterns for use in the training set. The EWFD was demonstrated in full-scale tests on the ex-USS SHADWELL.

Title: Validation of screening test kits for the determination of aflatoxins in nuts

Authors: E. Trullols[1], Itziar Ruis nchez[1], F.X. Rius[1], M. Odena [2], M.T. Feliu[2]

1 Departament de Quimica; Universitat Rovira i Virgili de Tarragona; Pla a Imperial Tarr co, 1; E43005 Tarragona, Catalonia; Spain

2 Public Health Laboratory; C/ M. Cristina n…54; E43002 Tarragona, Catalonia; Spain

Keywords: screening, test kit, validation, performance curve, aflatoxin

Presenter: Itziar Ruisanchez, ruisan@quimica.urv.es

In the last years, there has been an important trend of analytical developments towards fast screening methods and precise but easily applied techniques. Among them, screening test kits, commercial package containing all the reagents and sometimes the instrumentation for the analysis, has grown to a great extent [1]. The key point is that nowadays we are more interested in knowing whether or not the concentration of a specific analyte is above or below to a regulatory value than in quantifying a particular concentration. Therefore, screening systems have been developed to provide binary responses of the type 'yes/no', that are used for taking immediate decisions, for instance, whether the sample complies with the normative.

As it is well known, the ISO 17025 (previously, EN 45001) standard provides the basic guidelines for achieving the required level of quality in the analytical laboratories. Although a lot has been done in the establishment of the quality requirements for quantitative analytical methods, *e.g.* by the European Committee for Standardization or the Association of Official Analytical Chemists International, few improvements have been carried out with qualitative methods of analysis and, specifically, with screening analytical systems [2].

In this work we describe the performance characteristics of the commercial test kit Aflacard 2 ppb applied to the determination of aflatoxins in some nuts (pistachios, peanuts, etc.). Numerical values for performance characteristics like sensitivity, selectivity, false positive and false negative rate, uncertainty, etc. are established by means of the performance characteristics curves [3].

1) J. Stroka and E. Anklam. Trends in Anal Chem 2002;21:90-95

2) EURACHEM. The Fitness for Purpose of Analytical Methods, 1998

3) R. Song, P.C. Schlecht, K. Ashley. Journal of Hazardous Materials 2001;83:29-39.

Title: Monotonicity on identification with neural nets

Authors: Guillermo B. Sentoni, Lorenz T. Biegler
School of Engineering and Science; Universidad Argentina de la Empresa;
Lima 717; C1073AAO Buenos Aires; Argentina

Keywords: monotonicity, nonlinear identification, NNs

Presenter: Guillermo Sentoni, gsentoni@uade.edu.ar

This paper presents a training scheme to obtain a monotonic function approximation for a Single Hidden Layer Perceptron from sample data. A monotonic function is defined as having the property either of being never increasing or of being never decreasing as the values of the independent variable increase. If the function is multivariable with input dimension d, the definition still holds by input-output pairs: for example the input i changes while the other inputs are held at a constant value.

There exists cases in which monotonicity is a desirable property for an aproximator, specially in those applications in which a priori process knowledge is handy and well established. In those cases, process or system knowledge dictates that the relationship between some input/outputs pairs must be monotonically increasing or decreasing. Among other examples, models in Model Predictive control must fit the right static gain among inputs-outputs pairs. Achieving this requirement is not easy, even more, when sample data is drawn from a noisy process.

The objective of this training algorithm is to approximate a function through sample data while at the same time meeting given gains. The information provided to the algorithm are the training samples and a set of indexes that explain the gain of an input-output pair. Doing so, the user is able to specify if an input-output pair is monotonic descendent or monotonic ascendant, or runs free. The goal is to get a monotonic approximation of the underlying function represented by the data by constraining some of the weights of the neural network. Establishing the right constraints on the weights of the network will enforce monotonicity in a very elegant way. Results will be provided by testing the algorithm with data coming from a polymerization process with 11 inputs and 1 output.

Title: Automatic declutter approaches using GLS with discrete or continuous independent variables

Authors: Jeremy M. Shaver[1], Steven Wright[2]

1 Eigenvector Research, Inc.; PO Box 561; Manson, WA 98831; USA

2 Pioneer Hi-Bred International; PO Box 1004; Johnston, IA 50131; USA

Keywords: Generalized Least Squares, OSC, PCA, PLS

Presenter: Jeremy Shaver, shaver@eigenvector.com

Various preprocessing methods have been devised to remove variance that is orthogonal to the property of interest (predicted y-block variable) and all have shown some value in allowing simplified PLS models. Each, however, uses slightly different approaches and is more or less effective depending on the type and extent of systematic variations observed. In addition, each requires different levels of computational resources.

In this work, we show the application of three common orthogonalization or decluttering techniques to the prediction of isoflavone content in ground soybeans. In this example, the major spectral variations observed within the spectra arise from the normal variation of the major constituents (oil 15-25% and protein 30-45%) while isoflavone contributed only 0.1 to 0.5% of the spectral signal. That is, roughly 99.5-99.9% of the total X-block variance is orthogonal to the isoflavone variations. Several decluttering techniques were used in an effort to better understand the variance in the data and to investigate the ability of the techniques to stabilize model performance.

Title: Comparison of statistical confidence-limit estimates for sum-squared residuals

Authors: Jeremy M. Shaver, Barry M. Wise, Neal B. Gallagher
Eigenvector Research, Inc.; PO Box 561; Manson, WA 98831; USA

Keywords: residuals, confidence limits, chi-squared distribution

Presenter: Jeremy Shaver, shaver@eigenvector.com

In 1979 Jackson and Mudholkar [1] proposed a method to estimate confidence limits for sample sum of squared residuals (Q values) from principal components analysis models. The method was based on an analysis of the distribution of residual eigenvalues. It has been observed that this approach provides poor estimates of confidence limits for data sets with fewer samples than variables. An alternative method for estimating confidence limits on sum of squared residuals is based on fitting the degrees of freedom and scale factor to a Chi square distribution using the method of moments. Monte Carlo simulations were used to compare it to Jackson's method. Results showed that the Chi square approach provided estimates that agreed well with the data and provided better estimates for data sets with fewer samples than variables. Both approaches gave similar estimates for data sets with more samples than variables. It was also found that required computation time was considerably less for the Chi square approach. Comparisons are also presented for some real data sets.

1) J.E. Jackson, G.S. Mudholkdar. Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics 1979;21(3):341-349.

Title: Chemometrics: Regulatory challenges and opportunities in pharmaceutical manufacture

Authors: John A. Spencer[1], L.F. Buhse[1], M.M. Nasr[1], A.S. Hussain[2]

1 Division of Pharmaceutical Analysis; U.S. Food & Drug Administration; 1114 Market Street; St. Louis 63101; USA

2 Office of Pharmaceutical Science; U.S. Food & Drug Administration; Rockville, MD 20852; USA

Keywords: pharmaceutical manufacture, process analytical technology, drug quality, pharmaceutical approvals, FDA regulation

Presenter: John A. Spencer, spencerj@cder.fda.gov

In 2001 the FDA initiated a program to stimulate the development of Process Analytical Technology (PAT) in the manufacture of pharmaceuticals. This calls for substantial changes not only in the manufacturing technology but also in the philosophy of product quality control, the drug approval process and, ultimately, the regulatory atmosphere. Several meetings between interested parties — industry, academia and the FDA — have already been held to establish some basic practical guiding principles for use of PAT.

There are a variety of fast, non-destructive instruments and sensors such as near-infrared, Raman and particle sizing now in use by non-regulated industries. These will need to be implemented with detailed validation protocols that can be properly managed by the manufacturer and likewise understood by the FDA. Measurements from such devices will inevitably be joined into multidimensional process analytical-based control models.

Central to the implementation of PAT is the use of chemometric tools such as PCA and PLS. Choice and validation of control models will need to be appropriate to ensure product safety. Ultimately, wise use of chemometrics in the implementation of PAT will yield improved quality drugs, better regulated, more efficient processes and lower production costs resulting in benefits to the consumer, industry and the FDA. We will discuss some important issues that arise from the utilization of multivariate process control. Guidance will be needed for the review of FDA drug applications, for FDA compliance investigators and for pharmaceutical manufacturers. A FDA perspective on the impact of implementation of emerging process analytical technologies and associated chemometrics on drug quality will also be discussed.

Title: Simultaneous quantitation of HIV DNA vaccine constituents by second-derivative UV absorbance with multi-component analysis

Authors: Joyce A. Sweeney, Bettiann Waldner, Pei-Kuo Tsai
Merck & Co., Inc.; WP17-101, PO Box 4; West Point, PA 19486; USA

Keywords: HIV vaccine, UV absorbance, second-derivative, multi-component

Presenter: Joyce Sweeney, Joyce_Sweeney@Merck.com

Chemometrics is a powerful tool that permits direct analysis of test samples without tedious sample preparation steps (*e.g.*, extractions or chromatographies) by deconvoluting information from, in this case optical spectral data, using mathematical algorithms that are scientifically and statistically sound. The chemometric method reported here, employing second derivative UV absorbance with Classical Least Squares (CLS) analysis, was developed to simultaneously determine the DNA and excipient surfactant content in HIV DNA plasmid vaccine formulations containing a synthetic tri-block co-polymer, poloxamer CRL-1005, used as a vaccine adjuvant.

While the relative amounts of the various components in this HIV DNA vaccine result in highly desirable formulation properties, the complex matrix, rich in UV chromophores with overlapping spectral features, is not amenable to a simple UV concentration analysis for either DNA or the surfactant. The major obstacle for this simultaneous UV absorbance assessment is the relatively high concentration and high near-UV absorptivity of DNA compared to that of the surfactant in the vaccine. By exploiting the resolving power of second derivative absorbance spectroscopy and the higher molar absorptivity of the surfactant in the far-UV region at approximately 210 to 220 nm, where the relative absorbance contributions of DNA and surfactant at the formulation concentrations are more similar, it is possible to increase the sensitivity of the assay for the surfactant.

Absorptivities for DNA and surfactant are individually calculated throughout a specified wavelength region based on absorbance measurements of calibration standards. These absorptivities are used to calculate, in an iterative fashion, the best relative contributions of each component to an unknown multi-component sample measurement. The unique spectral lineshapes and relative intensities of each component ensure appropriate fits of the individual components. The sum of the calculated individual component spectra determined from fitting to an unknown multi-component sample measurement should ideally overlay the measured multi-component spectrum of that same unknown sample. Any discrepancy in the calculated multi-component spectrum relative to the measured multi-component spectrum represents an estimate of the fit error. An acceptance criterion for this method requires that the fit error be less than 5% for each component.

In addition to verifying highly acceptable levels of accuracy and precision for these determinations, method development included investigations to ensure the absence of any spectral shifts due to interactions of DNA, surfactant and/or adjuvant in multi-component formulations relative to single component calibrators. These latter investigations also verified that the variable lengths of an alkyl side-chain in the surfactant did not impact the molar absorptivity of the UV chromophore in the surfactant within the spectral region of interest, even in the presence of DNA and CRL-1005 poloxamer. The 2nd derivative UV absorbance method with CLS analysis is highly

efficient, quite rugged and generates very accurate and precise data.

Title: Background suppression strategies for the detection of volatile organic compounds by airborne passive Fourier-transform IR spectrometry

Authors: Toshiyasu Tarumi, Gary W. Small
Department of Chemistry & Biochemistry; Clippinger Labs; Ohio University; Athens, OH 45701-2979; USA

Keywords: passive FT-IR, VOCs, digital filtering

Presenter: Toshiyasu Tarumi , tt275388@ohio.edu

Constantly changing background radiation from the ground is an important factor that complicates the use of airborne passive Fourier transform infrared spectrometers for remote chemical vapor sensing. One way to solve this problem is to collect as much data as possible such that classification algorithms can be trained to account for the variety of background conditions encountered. However, this approach is not practical in terms of the cost and effort required for the data collection.

A more feasible solution is to suppress the background radiance with appropriate data processing methods. If the background effects can be eliminated by signal processing methods, it should be possible to build a classifier for the target analyte with laboratory data alone. In a series of studies, we have demonstrated that using short segments of digitally filtered interferograms can significantly reduce the effects of background differences. The background suppression in this case is based on differences in the decay rate of the interferogram signal. The interferogram representation of broader background spectral features decays much faster than that of narrower analyte spectra.

Spectra themselves can also be used as features for classification as long as the difference in the intensity and shape of the background radiation can be compensated. For example, the application of high-pass digital filters to single-beam spectra should minimize the broad spectral features of the background radiation.

This study focuses on the detection of ground sources of ethanol and methanol vapors with a downward-looking spectrometer mounted on an aircraft platform. Support vector machines trained with data collected on the ground are used to detect these volatile organic compounds. Comparisons will be made between the two background suppression strategies, interferogram processing and high-pass filtering of single-beam spectra. Infinite impulse response digital filters will be used in this work because of their sharp cutoff responses. Through these studies, the overall feasibility of training the classifier for the airborne application using ground data will be discussed.

Title: SpaRef: A new clustering algorithm for satellite imagery

Authors: Tran N. Thanh, R. Wehrem, L.M.C. Buydens
Laboratory of Analytical Chemistry; Katholieke Universiteit Nijmegen;
Toernooiveld 1; NL6525 ED Nijmegen; The Netherlands

Keywords: clustering algorithm, multispectral image segmentation, spatial analysis

Presenter: Tran N. Thanh, tnthanh@sci.kun.nl

Multispectral satellite images provide detailed data with information in both the spatial and spectral domains. Many segmentation methods for multispectral satellite images are based on a per-pixel classification that uses only spectral information and ignores spatial information. A clustering algorithm based on both spectral and spatial information would produce better results.

This work presents SpaRef, such a clustering algorithm for multispectral satellite images. Spatial information is integrated with partitional and agglomeration clustering processes. The number of clusters is automatically identified. SpaRef is compared with a set of well-known clustering methods. The clusters obtained show improved results.

Title: Example of a three-step strategy PCA-PLS-LDA with archaeometric data: Identification of an organic material on a Neolithic statuette

Authors: Kurt Varmuza[1], F. Sauter[2], W. Werther[3], P. Stadler[4]

1 Laboratory for Chemometrics; Institute of Chemical Engineering; Vienna University of Technology; Getreidemarkt 9/166; A1060 Vienna; Austria

2 Vienna University of Technology; Institute of Applied Synthetic Chemistry; Austria

3 University of Vienna; Institute of Analytical Chemistry; Austria

4 Museum of Natural History; Department of Prehistory; Austria

Keywords: EDA, classification, taxonomy, archaeometry

Presenter: Kurt Varmuza, kvarmuza@email.tuwien.ac.at

In an Early Neolithic settlement found near Vienna, Austria, pieces of terracotta figurines were found. Traces of an organic material on them were analyzed by means of GC/MS followed by multivariate data analysis as summarized below. This organic material could be identified as birch bark pitch, a material frequently used for many purposes in prehistoric Europe.

To identify the type of wood that have been used to prepare the ancient pitch, model pitches were prepared in the laboratory with material from four types of trees: betula, alnus, corylus, carpinus. In order to get rid of most of the atypical components the model pitches as well as the archaeological samples were Kugelrohr (bulb-to-bulb) distilled under reduced pressure (ca. 300¡C, 22-26 mbar), yielding a viscous oil that contained most of the significant terpene fraction. After purification by solid phase extraction this oil was analyzed by GC/MS. Each sample was characterized by the relative concentrations of 50 measured substances being mainly triterpenes. Data evaluation was performed as follows.

First step was principal component analysis (PCA). In the resulting scatter plot the model pitches formed partly overlapping clusters according to their origin. The archaeological sample was located within the cluster formed by birch samples (betula).

Second step was partial least-squares discriminant mapping (PLS) using the 50 concentrations as X matrix. The Y matrix contained four columns with the class information about the origin of the samples. The scatter plot of the first two PLS X-components showed an enhanced separation of the different families of trees with the archaeological sample clearly located in the betula area.

Third step was a linear discriminant analysis (LDA) to maximize the separation between birches and other tree families. The archaeological sample could be definitely assigned to the birch trees.

Title: Multivariate data analysis of chemical structure sets represented by binary substructure descriptors

Authors: Kurt Varmuza, H. Scsibrany, M. Karlovits, W. Demuth, F. M ller
Laboratory for Chemometrics; Institute of Chemical Engineering;
Vienna University of Technology; Getreidemarkt 9/166; A1060 Vienna; Austria

Keywords: cluster analysis of chemical structures, PCA, PLS, software SubMat

Presenter: Kurt Varmuza, kvarmuza@email.tuwien.ac.at

Among the many approaches for representing chemical structures as vectors the method using binary substructure descriptors plays an important role. Substructures are easily interpretable in terms of chemistry; they are capable to cover a great diversity of chemical structures, and substructure search is a standard operation in computer chemistry.

Software SubMat was developed for easy and automatic calculations of binary substructure descriptors for a set of molecular structures and a set of substructures (contained in two input files in Molfile format). SubMat generates an output text file with one line for each molecular structure containing a string of 0 s and 1 s for the absence or presence of the substructures. SubMat is running under MS Windows 95/98/2000/NT. Computing time for 1000 molecular structures and 100 substructures is typical 3 seconds on a PC with 1 GHz.

SubMat can be optionally executed by calling it from another program (for instance from a Matlab program). In this case a command file is used to transfer file names and parameters to SubMat. During execution semaphore files are used to communicate with the calling program to transfer for instance error messages, status data, or a stop command. In this remote mode no window is opened by SubMat.

A set of 1365 substructures has been defined that covers a wide area of organic chemistry; partly they have been obtained by applying the isomer generator software Molgen.

Sets with typical 20 to 200 molecular structures are for instance obtained from database searches or spectral similarity searches. PCA is a versatile tool for a cluster analysis of these chemical structures. If structural and spectral data are available PLS mapping is a powerful method to visualize or extract spectra-structure relationships. A number of other automatic or interactive software tools support the evaluation of such data. A KNN-type search can be applied to find substructures that are characteristic for compact clusters of molecular structures. A chi-squared test can be applied to extract structural differences between two clusters.

Acknowledgments to A. Kerber and R. Laue (University of Bayreuth, Germany) for providing the isomer generator software Molgen. Project P14792 of the Austrian Science Fund.

Title: Automated assays of radionuclides in chemical waste by passive gamma-ray spectroscopy and chemometrics

Authors: M. Elena Velasquez, Peter de B. Harrington, Ken Bosworth
Department of Chemistry & Biochemistry; Clippinger Labs; Ohio University; Athens, OH 45701-2979; USA

Keywords: wavelet, calibration, gamma-ray spectroscopy, homeland security, detection

Presenter: M. Elena Velasquez, Peter.Harrington@Ohio.edu

A fast and accurate method for determination of radioactive waste is important for storage and transport. Recent interest in detection of dirty bombs (*i.e.*, bombs that spread radioactive waste) and Homeland Security has increased the requirements for monitoring and inventory of nuclear waste. Passive gamma-ray spectra are acquired with portable energy dispersive detectors. These spectra allow *in situ* measurements of radionuclides. An automated intelligent system allows the detection and quantification of radionuclides. The computer based intelligent system provides real-time interpretation and avoids subjective biases that often lead to over-estimation of radionuclide concentration when interpreted by experts. Wavelet processing removes noise from the gamma spectra and reduces the dimensionality of the data sets. Multivariate calibration models were evaluated for estimating activity and mass of radionuclides.

Title: Algorithms for rapid detection of volatile organic compounds by passive Fourier-transform IR measurements from an aircraft platform

Authors: Boyong Wan, Gary W. Small
Department of Chemistry & Biochemistry; Clippinger Labs; Ohio University; Athens, OH 45701-2979; USA

Keywords: passive FT-IR, VOCs, pattern recognition, digital filtering

Presenter: Boyong Wan, wanby@yahoo.com

Fourier transform infrared (FT-IR) remote sensing measurements are used to implement an automated detection algorithm for volatile organic compounds (VOCs). Through the use of a combination of bandpass digital filtering and pattern classification techniques, the detection procedure can be performed directly on short segments of the interferogram data collected by the FT-IR spectrometer.

With optimized bandpass parameters, digital filters can extract specific frequencies associated with the spectral bands of the target analyte vapor, while eliminating the unwanted features arising from the infrared background or spectral interferences. Thus, a separate measurement of a background or reference spectrum can be avoided. Subsequently, pattern classification methods such as piecewise linear discriminant analysis or artificial neural networks are applied to the filtered interferogram segments to generate yes/no decisions regarding the presence of the analyte. When the spectrometer is mounted on a moving vehicle such as an aircraft, these processing steps allow the implementation of an automated, real-time analysis technique that allows a site to be quickly surveyed for the presence of a target vapor.

In the work presented here, methodology is developed for the detection of plumes of ethanol and methanol released from heated stacks and the influence of other potential interfering gases is investigated. A passive FT-IR spectrometer mounted on an aircraft platform is used to acquire the data in a downward-looking mode. A focus of the current work is the design of training protocols that will allow automated classifiers to be developed with laboratory or ground reference data and then to be subsequently applied to data collected from the air. In this process, several parameters, such as the composition of the training set, choice of classifier, filter position, filter width, filter stopband attenuation, and interferogram segment position must be optimized. This presentation will discuss the impact of these parameters on classification performance and will assess the strengths and weaknesses of airborne detection of VOCs.

Title: A multivariate approach to static ToF-SIMS image analysis of micropatterned surfaces

Authors: Bronwyn Wickes, David G. Castner
Department of Chemical Engineering; University of Washington;
NESAC/BIO, Box 351750; Seattle, WA 98195-1750; USA

Keywords: surface analysis, chemical state imaging, ToF-SIMS, PCA, micropattern

Presenter: Bronwyn Wickes, bronwynw@u.washington.edu

Novel biomaterial surfaces are being designed to specifically interact with their biomolecular environments. These surfaces may be patterned with multiple species of protein, peptide, and/or bio-inactive molecules to generate regions of differing bioactivity.

Static ToF-SIMS imaging offers a modality for simultaneously visualizing the spatial distribution of different surface species. Because ToF-SIMS imaging generates a full mass spectrum at each pixel, it is possible to use characteristic mass fragments to identify and differentiate between regions of different chemistry with a spatial resolution of approximately one micron. However, the utility of these datasets may be limited by their large size, degraded mass resolution and low ion counts per pixel.

Through spectral data denoising and multivariate image analysis, regions of similar chemistries may be more readily differentiated in ToF-SIMS image data. Micro-patterned surfaces were prepared with different species of proteins and thiol molecules. These surfaces were analyzed with ToF-SIMS imaging. Denoised image data was analyzed with principal component analysis to identify the combination of mass fragments that provided the best image contrast and chemical species identification.

Title: Resolution of HPLC peaks for Chinese medicines using wavelet analysis and evolving factor analysis

Authors: Lu Xu, Yu-Hua Qi
Changchun Institute of Applied Chemistry; Academia Cinica;
Changchun 130022, China

Keywords: Chinese medicines, HPLC, wavelet analysis ,EFA

Presenter: Lu Xu , luxu@ns.ciac.jl.cn

The compositions of Chinese medicines are usually very complicated. Cordyceps is a noble Chinese medicine that can be effectively used to several diseases. This medicine includes mainly nucleosides, sterols and polysaccharides. For quantitations of these compositions, the samples of cordyceps have been tested by using HPLC. The seriously overlapped spectra were resolved using wavelet analysis and evolving factor analysis with satisfactory results.

CAC - 2002

Best Poster Competition
Ballot for one poster

Poster selected

CAC - 2002

Best Poster Competition
Ballot for one poster

Poster selected