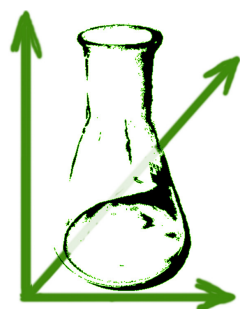


*The international meeting of chemometricians and statisticians
in Czech Republic*

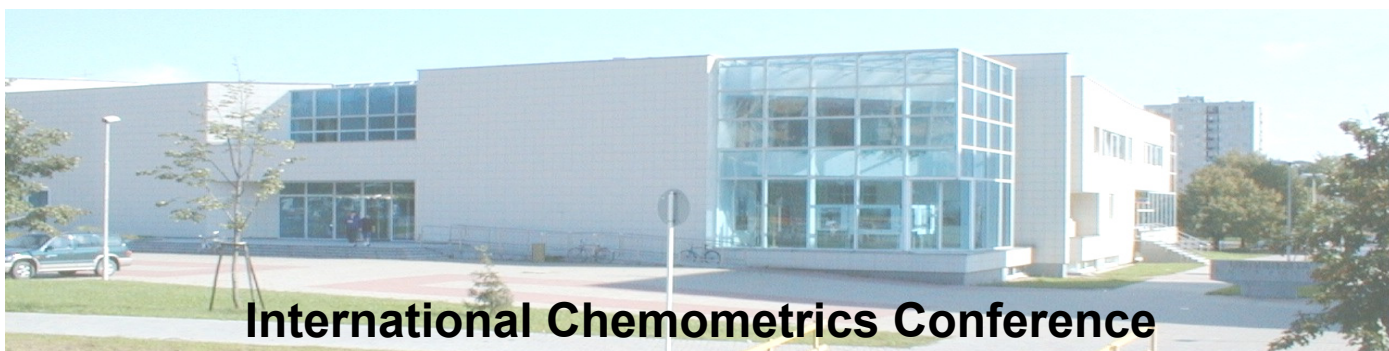


2004



ChemStat

Book of Abstracts



International Chemometrics Conference

August 30 to September 2, 2004
Pardubice, Czech Republic
a satellite meeting of Compstat 2004

CHEMSTAT 2004 organized by TriloByte Statistical Software, University of Pardubice, University of Liberec, Czech Chemical Society and its Chemometrics section.

© 2004 TriloByte

List of authors
(alphabetically)

| Surname | First name | Country | Paper Title |
|-------------------|-------------------|----------------|---|
| Nagwa | Abo El-Maali | Egypt | Simultaneous Voltammetric Chemometric Determination of the Anticancer Drugs: Tarabine PFS, Adriplastina and Methotexate |
| Evelyn | Archery | South Africa | Simultaneous Multicomponent Analysis of Cu, Ni, Fe and Co Using Partial Least Squares |
| Adegbola | Ayobami | Nigeria | Exploratory Data Analysis - Addressing the Context in which Chemometrics Works |
| Marek | Brabec | Czech Republic | State-Space Dynamic Model for Estimation of Radon Entry Rate, Based on Kalman Filtering |
| Richard | Brereton | United Kingdom | 1:Applications of Pattern Recognition; 2:Statistical Use of Analytical Chemical Data in Court Cases |
| Walter | Focke | South Africa | 1:Correlating Physical Property Data with Canonical Pade Approximants; 2:A Wilson Local Composition Model for the Viscosity of Liquid Mixtures |
| Jaroslava | Hálová | Czech Republic | EA of Structure-Activity Data (QSAR) using GUHA |
| Liudmila | Larina | Russia | The Analytical Signal Modelling for the Analysis Accuracy Increase at Some Model Determination by Stripping Voltametry in Environmental Objects |
| Milan | Meloun | Czech Republic | Minimizing the Effects of Multicollinearity in the Polynomial Regression |
| Harald | Martens | Denmark | Multivariate and Multiblock Chemometrics: A Culture for the "New Bioinformatics"? |
| David | Milde | Czech Republic | Teaching Chemometry and Good Laboratory Practice at Palacky University in Olomouc |
| Jiří | Milítký | Czech Republic | Statistical Modelling in Chemometrics |
| Muili | Mubarak Akorede | Nigeria | Development of Chemometric Software |
| Afshin | R. Khorrami | Iran | Determination of Caffeine in Blake Tea Leaves with FTIR Spectrometry Using Chemometrics A Comparison Between MLR, PLS, CLS as Calibration Models |
| Gbadamosi | Ramoni Adewale | Nigeria | How to Design Software |
| Selvarajah | Shankar Kumar | India | Wavelength Based Classification Technique |
| Sujatha | Subramanian | India | Optical Spectrum Analysis |
| Hana | Šormová | Czech Republic | Statistics and its Application During the Analysis of Optical Spectra; Uncertainties and Measured Spectroscopy Data |
| Tomáš | Syrový | Czech Republic | Number of Components Using Modified PCA Scree Plot in Spectroscopy |
| Michel | Tenenhaus | France | PLS and Sensory Analysis |
| Johan | Trygg | Sweden | Recent Developments in PLS Regression: OSC Filters, Pure Profile Estimation and Bi-directional (X-Y) Modeling |
| Vincenzo Esposito | Vinzi | Italy | PLS Generalised Linear Regression: Foundations and Recent Advances with Variable Selection and Validation Procedures |
| Zdeněk | Wagner | Czech Republic | Robust Estimation of Particle Size Distribution of Atmospheric Aerosols by Gnostic Theory |
| Manuel | Zarzo Castello | Spain | 1:Analysis of multiblock data sets to identify the key process variables in a batch polymerisation; 2:Consistency analysis for PLS with variable selection to diagnose a dehydration process; 3:Analysis of dynamic gas sensor response using chemometric methods |
| Simeone | Zomer | United Kingdom | Active Learning Support Vector Machines For Optimal Subset Selection in Classification |

Abstracts

PLS Generalised Linear Regression: foundations and recent advances with variable selection and validation procedures

Vincenzo Esposito Vinzi

Universita degli Studi di Napoli; vincenzo.espositovinzi@unina.it

PLS (Partial Least Squares) univariate regression (PLS1) is a model linking a numerical dependent variable to a set of numerical (or dummy) explanatory variables especially feasible in situations where multiple regression is unstable or not feasible at all (strong multicollinearity, small number of observations compared to the number of variables, missing data). The same kind of problems may be encountered also in classical logistic regression and, more generally, when using a generalised linear model. It is possible to apply the same principles of PLS regression to logistic regression as well as to generalised linear models.

PLS1 can be actually obtained by means of an iterated use of simple and multiple regressions based on ordinary least squares (OLS). By taking advantage from the statistical tests associated with linear regression, it is feasible to select the significant explanatory variables to include in PLS regression and to choose the number of PLS components to retain. The principle of the presented algorithm may be similarly used in order to yield an extension of PLS regression to PLS generalised linear regression (PLS-GLR). PLS generalised linear regression retains the rationale of PLS while the criterion optimised at each step is based on maximum likelihood. Nevertheless, the acronym PLS is kept as a reference to a general methodology for dealing with a set of predictors.

The approach proposed for PLS-GLR is simple and easy to implement. Moreover, it can be easily generalised to any model that is linear at the level of the explanatory variables. Some examples show the use of the proposed methods in real practice with specific reference to classical PLS regression, PLS logistic regression and the application of PLS-GLR.

PLS and Sensory Analysis

M. Tenenhaus¹, J. Pages², L. Ambroisine³, C. Guinot^{3}*

¹HEC School of Management-Paris, France; ²ENSA-INSFA, France; ³CE.R.I.E.S., France

tenenhaus@hec.fr

In this paper a new methodology devoted to the analysis of a type of data often encountered in sensory analysis is described. We are interested in daily used products like orange juices, yogurts, lipsticks. A small number of products are described by physico-chemical and sensory characteristics. Moreover these products are evaluated by consumers on a preference scale. The objective of the statistical analysis is to relate the hedonic block of variables Y (the consumers' preferences) to the physico-chemical block X_1 and to the sensory block X_2 . Generally the data table to be analyzed has about 10 rows, about 30 predictors X , and about 100 responses Y . Some data can also be missing. PLS methods are perfectly suitable for this type of problem. PLS regression allows the analysis of the link between the responses Y and the predictors $X = [X_1, X_2]$. Using this method it is possible to cluster the consumers in homogeneous groups with respect to their tastes and such that their behavior can be related to the characteristics of the products. For each homogeneous group of consumers PLS regression allows to obtain a graphical display of the products with their characteristics and a mapping of the consumers based on their preferences, or contour lines on product preferences. Statistical validation can be obtained by Jack-knife. PLS path modeling allows a more detailed analysis of each homogeneous group of consumers by building a causal scheme: each homogeneous block of consumers is related to the physico-chemical block X_1 and the sensory block X_2 , and the sensory block is itself related to the physico-chemical block. Statistical validation is carried out by Bootstrap. PLS methods allow to obtain more complete and of easier understanding than usual methods: PLS methods fit sensory analysis like hand in glove.

Keywords: PLS regression, PLS approach, PLS path modeling.

Recent developments in PLS regression: OSC filters, pure profile estimation and bi-directional (X-Y) modeling

Johan Trygg, Svante Wold

Research group for Chemometrics, Institute of Chemistry, Umeå University, Sweden

The original chemometrics PLS model [Ref.1] with two blocks of variables (\mathbf{X} and \mathbf{Y}) has seen several modifications since the beginning of 1980. Often, the \mathbf{X} matrix represents spectral or chromatographical profiles and the \mathbf{Y} matrix contains the corresponding concentration profiles of the known analytes in \mathbf{X} . It is also possible to infer other properties \mathbf{Y} of samples, e.g. the strength of polymers, or even complementary spectral / chromatographical measurements.

Lately, the understanding of the “mechanisms” of PLS estimation has increased substantially. We now know that PLS (and similar methods) are affected by the presence of systematic variation in \mathbf{X} that is not related to \mathbf{Y} , here defined as *structured noise*. Examples of structured noise are the variation of baseline, unknown constituents, or effects of changing instrumental equipment. Structured noise increases the number of PLS components for each “latent direction” in \mathbf{Y} , and complicates model interpretation [Ref. 2].

Recently, Wold et al. developed the orthogonal signal correction (OSC) method [Ref. 3] to filter out structured noise from the \mathbf{X} matrix. This was followed by the O-PLS and O2-PLS methods [Ref. 2,4]. They underline that there exists only one Y-related component for a single Y-variable. In addition, besides providing good predictions, these methods are also capable of pure profile estimation [Ref. 4]. This has previously been regarded as an exclusive feature of the direct calibration methods such as classical least squares. The O2-PLS method [Ref. 4] can also model \mathbf{X} and \mathbf{Y} in a bi-directional fashion ($\mathbf{X} \rightarrow \leftarrow \mathbf{Y}$), unlike PLS. This provides the ability to find common and unique variations in large multiple data tables.

Two examples will demonstrate how these new features facilitate interpretation of complex multivariate systems. First, a multivariate calibration example with unknown constituents in the \mathbf{X} matrix is presented. The second example is from the area of functional genomics where multiple data tables containing mass spectral measurements (GC/MS) on biofluids were analysed.

References

1. Wold S, Ruhe A, Wold H, Dunn III W J. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. in SIAM J. Sci. Statist. Comput. 1984; 5:735-743
2. Trygg J, Wold S, Orthogonal projections to latent structures, O-PLS, J.Chemometr., 2002; 16: 119-128.
3. Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. Chemomet. Intell. Lab. Syst., 1998; 44: 175-185.
4. Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter, J. Chemometr., 2003; 17: 53-64.

Correlating mixture properties with homogenous Scheffé polynomials and Padé approximants

W. W. Focke and B. Du Plessis

Univ. Pretoria, Dep. Chemical Engineering, Lynnwood Rd., Pretoria 0002, South Africa

Tel. +27 12 420 2588, fax: +27 12 420 2516, e-mail: walter.focke@up.ac.za

Padé approximants have previously been limited to fitting physical property data for pure components or binary mixtures. Here we present revised canonical forms for homogeneous Scheffé polynomials and Padé approximants suitable for the modelling of multicomponent mixture properties. The general expression for a q -component homogeneous N^{th} order Scheffé K-polynomial is:

$$K_q(N) = \sum_{i \leq j \leq k \dots}^q M_{ijk\dots} c_{ijk\dots} x_i x_j x_k \dots \quad (1)$$

The $c_{ijk\dots}$ are adjustable parameters and the $M_{ijk\dots}$ are multinomial coefficients defined by:

$$M_{ijk\dots} = \binom{N!}{n_1! n_2! \dots n_q!} \quad (2)$$

Here $N = \sum_{k=1}^q n_k$ and n_k denotes the number of times the label k occurs in the subscript of the multinomial $M_{ijk\dots}$ or parameter $c_{ijk\dots}$.

The canonical Padé approximants are defined as the ratio of two canonical Scheffé K-polynomials:

$$P_q(M, N) \equiv \frac{K_q(M)}{K_q(N)} \quad (3)$$

This choice offers attractive symmetries, a compact notation and also satisfactorily correlates multicomponent physical property data.

Keywords: Rational function; modelling; mixture; physical property

Analysis of multiblock data sets to identify the key process variables in a batch polymerisation

Manuel Zarzo, Alberto Ferrer

Department of Applied Statistics, Operations Research and Quality

Polytechnic University of Valencia, Camino de Vera s/n ; edificio I-3 ; 46022 Valencia (Spain), mazarcas@eio.upv.es, aferrer@eio.upv.es

An industrial batch polymerisation has been studied. Every minute, 38 process variables (temperatures, pressures, flows, etc.) are recorded on line, and one of the quality parameters analysed in laboratory in the final product is the hydroxyl number (IOH). The problem is that this parameter is out of specifications in a certain amount of batches. To diagnose the causes, data from 69 batches (produced in two different periods) have been analysed. As the duration of the different stages is not constant from batch to batch, several alignment methods have been used to synchronise the trajectories. Each aligned "trajectory" is a matrix formed by the evolution of the corresponding process variable in a scale of pseudo-time for the set of 69 batches. The final matrix is a multiblock data set comprised by different trajectories, and batches from two periods. With every block of variables that comprises a trajectory, two different analyses have been conducted: a PCA and a PLS regression considering the IOH as response variable. In both models, all significant latent components according to cross-validation have been obtained. With this procedure, the initial matrix of 8985 variables is transformed into a matrix with 516 latent variables that contains nearly the same information. Afterwards, the squared linear correlation coefficient between every latent variable and the IOH has been calculated in three cases: with all batches, selecting those from the first period, and with the batches from the second one. Analysing the results, the correlation with IOH is statistically significant in 7 latent variables considering all batches, and also if each subset of batches is selected. But there are some latent variables where the correlation is significant only in one set of batches, reflecting that the causes of variation of this quality parameter might not be the same in both periods. These results focus on the key process variables that might require a better control to avoid batches out of specifications. Nevertheless, a design of experiments should be carried out to finally identify a cause-effect relationship with the hydroxyl number.

State-Space Dynamical Model for Estimation of Radon Entry Rate, Based on Kalman Filtering

M.Brabec¹, K.Jilek²

1 National Institute of Public Health, Department of Biostatistics, Praha, Czech Republic

2 National Radiation Protection Institute, Praha, Czech Republic

In this paper, we will describe a new approach to estimation of radon entry rate (RER) from a tracer gas experiment during which the tracer (CO) and radon concentrations are measured simultaneously. Such an experiment can be easily conducted e.g. in an ordinary building where one wants to know the RER, e.g. for monitoring purposes, judging whether preventive measures (like insulation) are necessary, or to assess their effectiveness. Since the Rn concentration is influenced both by the RER we wish to estimate, and by the air ventilation rate (AVR), it is not possible to obtain the estimate from Rn measurements only, without very strong assumptions which are hardly satisfied in practice. Synchronized tracer gas measurement helps to estimate AVR. Simultaneous measurement effectively leads to decoupling the RER and AVR effects. In order to do this, one essentially has to solve a pair of differential equations coming from a simple physical model, using measured data. Several ad hoc techniques have been used for this purpose in the past. Main disadvantage comes from their ad hoc and heuristic character. Specifically, they do not acknowledge presence of the measurement error and they are not derived from any explicitly stated formal statistical model, so that their performance is unstable, varies from bad to good, depending on circumstances whose effect is hard to foresee. Moreover, most of these methods rely on assumptions that are rather restrictive even in physical sense (steady state, etc.).

Our approach tries to overcome these difficulties by formulating an explicit dynamical statistical model. The model is of state-space type and hence it acknowledges the measurement error presence explicitly. Its' formulation rests in writing essentially a discrete-time analogue of the differential equations from the underlying physical model. Moreover, it allows for a rather general measurement variance structure to cover both homo- and hetero-scedastic cases (which is useful e.g. when dealing with radioactive counters having Poisson-like behavior). A particular parametrization is carefully chosen to retain a natural meaning of the state variables. The model formulation is modular and hence it offers flexibility and both theoretical and practical benefits. Even though the resulting representation is nonlinear in the states, state estimation can be handled relatively easily by means of Kalman filtering (particularly by the extended Kalman filter obtained by local linearization). Recursive character of the algorithm allows for easy on-line computations of both filter and prediction type for rather large datasets from longer measurement campaigns using detailed time resolution. Additional difficulty arises from the fact that the state-space model used contains several structural parameters (e.g. measurement error and structural disturbance variances) which are not known a priori. Their values have to be supplied in order to be able to run the filter. To overcome this problem, we use the maximum likelihood estimation procedure to estimate them. Fortunately, Kalman filter is extremely helpful for efficient (log)likelihood computation (thanks to the so called prediction error decomposition). After obtaining the MLE's, we run the filter with the structural parameter estimates plugged-in. Whole algorithm can be easily programmed in a modular way, which helps its implementation e.g. for various modifications of tracer gas experimental design. Apart from the description of the model, estimation procedure and discussion of their properties, we will illustrate their behavior on real-life data obtained from a prolonged measurement campaign conducted by NRPI.

KL Miner-A Tool of Data Mining for Patterns

Validation on Biological Activity Data

Jaroslava Hálová^a, Miloš Macháček^b and Jan Rauch^c

a Academy of Sciences of Czech Republic, Institute of Inorganic Chemistry, CZ 250 68 Rez

b Charles University, Faculty of Pharmacy, Heyrovského 1203, CZ 500 05 Hradec Kralove

c University of Economics, Winston Churchill Square 4, 130 00 Prague 3

The data mining procedure KL-Miner mines for patterns of the form $R \approx C / \gamma$. Here R and C are categorical attributes and γ is a Boolean attribute. The intuitive meaning of the expression $R \approx C / \gamma$ is that the attributes R and C are related in the way given by the symbol \approx when the condition given by the Boolean attribute γ is satisfied. The KL-Miner procedure deals with data matrices. The attributes R and C correspond to columns of the analysed data matrix and the Boolean attribute γ is derived from other columns of the analysed data matrix.

The symbol \approx is called KL-quantifier and it corresponds to a condition imposed by the user on the contingency table of R and C. The first experience [1] shows that the KL-Miner works is able to solve lot of practically important tasks in reasonable time.

In our analysis we used susceptibility data of *Mycobacterium tuberculosis*, *M. avium*, *M. kansasii*, and one clinical isolate to a series of 66 *N*-benzylsalicylamides substituted in both acyl and amide moieties [2]. Two types of analyses have been performed. First, minimum inhibitory concentrations after 14 days and 21 days were compared. Furthermore, structure-antimycobacterial activity relationships have been studied.

We used in our analysis the condition concerning Kendall's coefficient. This way it is possible to express ordinal dependences among values R and C. Linear dependencies of the biological activities after 14 and 21 days have been found – the application of KL-Miner to the analysis of biochemical data has been validated in this way KL-Miner represents an alternative to the classical data correlation analysis as a special case KL-Miner software system is a powerful tool of data analysis. Data mining in screening databases of pharmaceutical companies can be performed in this way. The developed methodology is a quite general data mining method. As a sequel KL-Miner is widely applicable beyond the scope of biochemical data analysis

References

- [1] Rauch J., Šimůnek M., Lín V.: Mining for Patterns Based on Contingency Tables by KL-Miner, First Experience. In: Lin T. Y., Hu X., Ohsuga S., Liao C. J. (ed.). Data mining – Foundations and New Directions in Data Mining. IEEE Computer Society, 2003, s. 156–163.
- [2] Waisser K., Peřina M., Klimešová V., Kaustová J.: On the Relationships between the Structure and Antimycobacterial Activity of Substituted *N*-Benzylsalicylamides. Collect. Czech. Chem. Commun. 68, 2003, 1275-1294.

Applications of Pattern Recognition

Richard G Brereton

School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, United Kingdom; r.g.Brereton@bris.ac.uk

Pattern recognition techniques in chemometrics have evolved considerably over the past few years. The lecture will illustrate both novel techniques and applications.

Most methods of chemometric pattern recognition involve the handling of analytical data which may be of numerous forms, for example GCMS peak areas, spectra, capillary electropherograms, DNA microarrays, and mass spectra. The crucial first steps involve preparing this data. GCMS is used to illustrate this stage. Peaks have to be identified, aligned in different GCMSs, quantified often involving baseline correction, smoothing and sometimes deconvolution, scaled (often there are significant problems relating to normalisation), and then the most useful peaks selected for subsequent pattern recognition steps. Each step has the potential for errors, and often the performance of pattern recognition algorithms depends crucially on these early steps.

Once a datamatrix has been obtained, for example of peak intensities for each sample, there are a significant number of pattern recognition methods to choose from. These include exploratory approaches such as principal components analysis. Classification using one class classifiers can be performed in three main ways. Modelling methods such as discriminant PLS are most familiar to the analytical chemist. Boundary methods including support vector machines are of increasing interest. Finally density methods can be applied. Different approaches also have an influence on the most appropriate size of training sets required, for example in the case of support vector machines it is often possible to determine adequate class models with a relatively small number of support vectors that fall on the boundaries, whereas for many modelling methods quite significant sample datasets are required to provide a suitable class model.

The most useful approach to pattern recognition depends very much on the type of problem under consideration and it is not possible to generalise. A range of applications are discussed. These include using GCMS in mammalian scent marking (e.g. how badgers recognise sex from urine); medical diagnosis using capillary electrophoresis; tumor classification from DNA microarrays; detection of environmental pollution using headspace mass spectrometry; determination of origins of pharmaceutical tablets using pyrolysis GCMS; genotyping in nutritional screening; and materials analysis using thermal profiling.

Statistical Use of Analytical Chemical Data in Court Cases

Richard G Brereton,

School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, *United Kingdom*

r.g.Brereton@bris.ac.uk

Analytical chemical data is frequently employed in forensic science.

Often the results of analytical tests are used to demonstrate guilt or innocence in court cases. These can be formulated mathematically in terms of obtaining probabilities of guilt. Juries often cannot understand detailed statistical analysis, so it is the job of the expert to provide guidance. If probabilities of guilt are above a certain level this is strong evidence for the prosecution.

In any court case, there are several pieces to the evidence, to each of which a probability can be attached. Bayesian statistics can be employed to come to a final verdict. This converts a prior probability (for example of guilt) to a posterior probability after the new evidence is included.

In the specific application, contamination of banknotes by drugs is measured by mass spectrometry. The higher this contamination, the higher the probability that the banknotes originate from a drug dealer. However there is still a small possibility that the banknotes originate from innocent sources, because contaminated banknotes will be spent by drug dealers and enter into circulation, also contamination rubs off from one banknote to another. The more banknotes analysed, the smaller the chance of a mis-conviction.

Likelihood ratios are often used in clinical chemistry to determine the probability of a disease if a test proves positive. These can also be applied in forensic science, so it is possible to determine a likelihood ratio that a defendant is guilty if the level of contamination of banknotes is above a certain threshold. This can be converted into a probability. It is then possible to employ Bayesian statistics to determine the number of batches of contaminated banknotes that are required to demonstrate probabilities of guilt. This also allows the analytical chemist to assess how useful a particular test is, and how many tests are required to produce evidence of a specific quality.

The analytical signal modelling for the analysis accuracy increase at some metals determination in environmental objects by stripping voltammetry

Romanenko S.V., Larina L.N., Larin S.V.

Tomsk Polytechnic University, Chemical Technology Department; lucy@anchem.chtd.tpu.ru

The analysis accuracy increase is one of the most acute problems in analytical chemistry. The determination of heavy metals in the environmental objects is complicated by its low concentrations in environment that lead to the insufficient precision of analysis. At determination of mercury(II), copper(II), lead(II), antimony(III) and bismuth(III) by stripping voltammetry the analytical signals seem negligible in comparison with the residual current. In this case at processing and interpretation of analytical signals of these metals the systematic error can be brought into results.

In most cases the correct estimation of the systematic error of analytical chemistry methods is complicated by the random error, conditioned by irreproducibility of some experimental factors. In stripping voltammetry the appreciable systematic error is connected with the non-linear base line subtraction

In our recent papers the technique of systematic error estimation by modelling of the analytical signal series with the empirical functions has been reported.

In this paper the dependences of systematic error on relative magnitude of analytical signal of some metals are plotted. The mean value of systematic error and its confidential interval are obtained and the experimental results are corrected.

The efficiency of the proposed technique is proved by determination of mercury(II), copper(II), lead(II), antimony(III) and bismuth(III) in the modelled solutions and environment objects that are samples of natural waters and soils of Tomsk region (Russia). It is shown, that after correction of the experimental results the essential approach of result value to the entered concentration is obtained. The confidential interval becomes significantly narrow (approximately in 2–3 times), that makes the proposed technique applicable for the correct determination of the low concentration of these metals in environmental objects.

To sum up, the proposed procedure of systematic error compensation allows reducing the summary error and increasing the analysis accuracy of determination of mercury(II), copper(II), lead(II), antimony(III) and bismuth(III) in environmental objects by stripping voltammetry.

Multivariate and multiblock chemometrics: A culture for the “new bioinformatics” ?

Harald Martens^{1,2}

¹*Norwegian Food Research Institute,*

²*Centre for Integrative Genetics, Norwegian Agricultural University*

Modern genomics and its associated fields of proteomics, metabonomics etc generate vast numbers of data that, in the end, need to be interpreted by human beings. A limiting factor is how to find, visualize and interpret reliable patterns of co-variation between different types of “-omics” data, in light of qualitative and quantitative background knowledge. Some molecular geneticists, biochemists and biologists working in the new “-omics”-fields feel alienated with traditional “school statistics” - both with the classical methods and with the academic statistical culture.

The field of Chemometrics has, over the last couple of decades, developed some powerful data analytical techniques and a pragmatic data analytical culture,- both of which may be useful in the “new bioinformatics”.

In particular, the versatility of cross-validated, graphically oriented Partial Least Squares Regression (PLSR) will be discussed. Then, some recent multi-block extensions of the PLS Regression (PLSR), suitable for interpreting more or less incompatible data tables, will be presented. They will be applied in various data sets to relate gene and gene expression data to proteome & metabonome data and to high-resolution phenotype characterization (spectroscopy) , and to background knowledge.

Two-block PLS: This well-established method relates $\mathbf{Y}(N \times J)$ to matrix $\mathbf{X}(N \times K)$ by extracting X-weights as left-hand singular vectors of $\mathbf{X}^T\mathbf{Y}(K \times J)$, either simultaneously or sequentially. This estimation criterion may be used in several different ways, yielding e.g. PLS Regression, Bookstein PLS analysis or Bridge-PLSR.

Three-block PLS: The *L-PLSR* relates $\mathbf{Y}(N \times J)$ to matrices $\mathbf{X}(N \times K)$ and $\mathbf{Z}(J \times L)$ by extracting X- and Z-weights as left- and right-hand singular vectors of $\mathbf{X}^T\mathbf{Y}\mathbf{Z}(K \times L)$, either simultaneously or sequentially. All rows and columns in \mathbf{X} , \mathbf{Y} and \mathbf{Z} may then be related graphically to the resulting latent variable structure in so-called correlation loading plots. This “endo-LPLSR” allows \mathbf{X} and \mathbf{Z} to be related to each other via \mathbf{Y} , even though they share no common dimension. Like conventional two-block PLSR this estimation criterion may be used in several different ways.

Four-block PLS: A new method, *U-PLSR*, will be presented. It is an extension of the three-block L-PLSR, extracting bi-linear structures in the four matrices $\mathbf{X}_1(N \times K)$, $\mathbf{Y}_1(N \times J)$, $\mathbf{X}_2(L \times K)$

and $\mathbf{Y}_2(M \times J)$ by svd of their coupled covariances, $\mathbf{X}_2 \mathbf{X}_1^T \mathbf{Y}_1 \mathbf{Y}_2^T$ ($L \times M$), either simultaneously or sequentially. This “endo-UPLSR” allows \mathbf{X}_2 and \mathbf{Y}_2 to be related to each other, via \mathbf{X}_1 vs \mathbf{Y}_1 , even though \mathbf{X}_2 and \mathbf{Y}_2 share no common dimension. Like conventional two-block PLSR this estimation criterion may be used in several different ways.

The U-PLSR represents a simple version of multiblock PLS network modelling, which may be useful for analysing complex biological systems under static and dynamic conditions.

ACTIVE LEARNING SUPPORT VECTOR MACHINES FOR OPTIMAL SAMPLE SELECTION IN CLASSIFICATION WITH APPLICATION IN MASS SPECTROMETRY

Zomer S.^a, Sánchez MDN.^b, Brereton RG.^a

a) Centre for Chemometrics, School of Chemistry, University of Bristol,
Cantock's Close, Bristol, BS8 1TS, UK.

b) Departamento de Química Analítica, Nutrición y Bromatología, Facultad de
Ciencias Químicas, Universidad de Salamanca, 37008 Salamanca, Spain.

Labelling samples in the analytical chemical laboratory is a procedure that may result in significant delays particularly when dealing with larger datasets and/or when labelling implies prolonged analysis. In such cases, a strategy that allows the construction of a reliable classifier on the basis of a minimal sized training set by labelling a minor fraction of samples can be of advantage. Support vector machines (SVM) are ideal for such an approach because the classifier relies on only a small subset of samples, namely the support vectors, while being independent from the remaining ones that typically form the majority of the dataset. This contribution describes a procedure where a SVM classifier is constructed with support vectors systematically retrieved from the pool of unlabelled samples. The procedure is termed as "active" because the algorithm interacts with the samples prior to their labelling rather than waiting passively for the input.

The learning behaviour on simulated datasets is analysed and a practical application for the detection of hydrocarbons in soils using mass spectrometry is described. Results on simulations show that active learning SVM optimally performs on datasets where the classes display an intermediate level of separation. On mass spectral data the classifier correctly assesses the membership of all samples in the original dataset by requiring for labelling around 14% of the samples. Its subsequent application on a second mass spectral dataset also provides perfect classification without further labelling, giving the same outcome of most classic techniques that were based on the entirely labelled original dataset.

Keywords: support vector machines, mass spectrometry, active learning.

E-mail: s.zomer@bristol.ac.uk

References

- [1] Pérez Pavón JL. et al. (2003). *A Method for the Detection of Hydrocarbon Pollution in Soils by Headspace Mass Spectrometry and Pattern Recognition Techniques*. Analytical Chemistry. **75**: 2034-2041.
- [2] Zomer S. et al. (2004). *Support Vector Machines for the Discrimination of Analytical Chemical Data: Application to the Determination of Tablet Production by Pyrolysis-Gas Chromatography Mass Spectrometry*. Analyst: **129**: 175-181.
- [3] *Proceedings of the 17th International Conference on Machine Learning*. (2000). Morgan Kaufmann Publishers, CA, USA: 111-118, 839-846, 999-1006.
- [4] Campbell C. (2002). *Kernel Methods: a Survey of Current Techniques*. Neurocomputing **48**: 63-84.

Robust estimation of particle size distribution in atmospheric aerosols by gnostic theory

Zdeněk Wagner¹, Vladimír Ždímal², Jiří Smolík²

¹E. Hála Laboratory of Thermodynamics, Institute of Chemical Process Fundamentals AS CR, Rozvojová 135, CZ-16502 Prague, Czech Republic.

²Aerosol Laboratory, Institute of Chemical Process Fundamentals AS CR, Rozvojová 135, CZ-16502 Prague, Czech Republic.

In recent years the motivation to study atmospheric aerosols increases due to the impact of aerosols on earth radiation balance. The aim is to estimate both the total budget of aerosol particles in the different parts of the atmosphere, their major sources and sinks, and also chemical composition. Still poorly understood are e. g. so called nucleation bursts, during which huge amounts of new particles are formed in the troposphere during a short time (Clarke, 1992). In order to follow aerosol dynamics of such events, continuous measurements of particle size distribution of atmospheric aerosol are necessary.

Such measurements are usually carried out using a combination of a differential mobility analyzer (DMA) with some suitable condensation particle counter (CPC), either in a differential (Differential Mobility Particle Sizer—DMPS), or a scanning mode (Scanning Mobility Particle Sizer—SMPS). After inverting the raw data the particle size distribution of the ambient aerosol is received (Kousaka et al., 1985). In addition, an aerodynamic particle sizer (APS) is sometimes used to extend the measured distribution towards larger particle sizes (Morawska et al., 1999).

The resulting distributions are usually multimodal. The reason is that each particle source produces particles with a unique size distribution and the atmosphere contains a mixture of particles coming from all these sources. As the atmosphere is a dynamical system with many simultaneous processes going on, these distributions are continuously changing with time. In order to minimize the amount of information for further data analysis, the distributions are often described using some characteristic values: mode positions, concentrations contained in each mode, etc. Several statistical approaches have been used so far, e. g. methods using moments (Barrett and Jheeta, 1996; Wright et al., 2002). Very often the assumption has been made that the resulting distribution can be represented by a sum of lognormal or gamma distributions, and parameters of these have been sought. However, as has been mentioned in the literature (Hinds, 1999), such a representation is justified more by practical reasons than by the underlying physics. Moreover, determination of correct number of modes by pure statistical method is a difficult task. J.S. Marron et al. (Marron, 1999) have developed a graphical tool which is freely available from the Internet (SiZer). However, the SMPS device can measure up

to 1000 distributions a day. Their routine analysis by means of a graphical interactive tool is practically impossible. The need for a robust automatic tool is thus urgent.

This work makes use of gnostic theory of uncertain data (Kovanic 1986, 1990, 2002). The gnostic kernel, which is derived from theory, is used for kernel estimation of the distribution function. The key problem is estimation of scale parameter which determines the width of the kernels. The method was tested by analyzing artificial distributions formed by linear combination of lognormal distributions disturbed by random errors with normal and Cauchy distributions. The method was then applied to real data from the SMPS. It was found that there exists a criterion for estimating an optimum value of the scale parameter which leads to the best description of the experimental data.

The method is built upon Octave which is a mathematical package freely available for all operating systems (www.octave.org). The gnostic analyzer, which is being developed in this work, will be available as free software in the future.

Acknowledgment

Support of this work by the GA ASCR under a grant No. IAA2076203, and by the GA CR under grant 205/03/1560 is gratefully acknowledged.

References

- [Clarke, 1992] Clarke A. D. (1992). Atmospheric Nuclei in the Remote Free Troposphere. *J. Atmos. Chem.* 14:479–488.
- [Hinds, 1999] Hinds W. C. (1998) *Aerosol Technology*. J. Wiley & Sons, New York, 2nd ed., p. 105.
- [Kousaka et al., 1985] Kousaka Y., Okuyama K., Adachi M. (1985). Determination of Particle Size Distribution of Ultra-fine Aerosols Using a Differential Mobility Analyzer. *Aerosol Sci. Technol.*, 4:209–225.
- [Kovanic, 1986] Kovanic P. (1986). A New Theoretical and Algorithmical Basis for Estimation, Identification and Control. *Automatica*, 22:657–674.
- [Kovanic (1990)] Kovanic P. (1990). Gnostická teorie neurčitých dat (Gnostic Theory of Uncertain Data). DrSc. Thesis, Prague.
- [Kovanic (2002)] Kovanic P. (2002). Private communication.
- [Marron (1999)] Chaudhuri P., Marron J. S. (1999). *JASA*, 94:807-823
- [Morawska et al., 1999] Morawska L., Thomas S., Jamriska M., Johnson G. (1999). The Modality of Particle Size Distributions of Environmental Aerosols. *Atmospheric Environment* 33:4401–4411.
- [SiZer] http://www.stat.unc.edu/faculty/marron/DataAnalyses/SiZer_Intro.html

Uncertainties and Measured Spectroscopy Data

Hana Šormová

*Institute of Physical and Applied Chemistry, Brno University of Technology, Purkyňova
118, 612 00 Brno, Czech Republic, hanka@milansorm.cz*

Uncertainties should be taken like the shadow of the measured data and include the statistic to their analysis. Values without uncertainties are only for the orientation, not exact.

Problems with the uncertainties determination and their applications to the measured data are not trivial. The methods must be analysed in their details and all of the aspects having possible influence it is necessary to include. The errors appeared during the spectra recording and their combinations are the most important uncertainty sources. So it is important to do the detail analysis which is only one way to the right interpretation of measured data.

Currently there exist the new concept of the uncertainty determination which disparts uncertainties in two categories (see [1–3]).

The calculation of the first type, called uncertainty A (u_{Ay}), is based on the statistic computation and the principle of the second type, called uncertainty B (u_{By}), is in the other determination expect the statistics. This type is usually caused by experiences of the user with the method, errors of apparatus, uncertainty of the measuring method, simplify the formula of the value calculation, approximation, extrapolation, non-compensative conditions of surroundings and so on. The definitive value of this uncertainty is characterised by the mean-root-square error and it can be determined for p sources by the following formula (1).

$$u_{By} = \sqrt{\sum_{j=1}^p A_j^2 \cdot u_{Bzj}^2}, \quad (1)$$

where A_j represents the factor of sensibility and u_{Bzj} are uncertainties of the sources.

The combination between the uncertainty A and B is identified like the combined standard uncertainty u_C and it is calculated according the formula (2).

$$u_C = \sqrt{u_{Ay}^2 \cdot u_{By}^2} \quad (2)$$

The combined standard uncertainty presents the interval where lie the real magnitude with the 68 % probability. That's why it is necessary to use the spread coefficient k (formula (3)) which increases the value of probability. If the k is equal 2, the probability is 95 %.

$$U = k \cdot u_C \quad (3)$$

This principle is general and it can be applied to measured optical spectra. And by this way it is possible to determine the accuracy of the measured data and calculated parameters of the experimental spectra. There is an example of the rotational temperature's computation and also of this theory's interpretation (see [5–7]). For this example was used the software program Simul (see [8]) which has been created for this situation. The obtained results are in the table 1.

| Magnitude (X_i) | Estimation of x_i | Standard Uncertainty u_{x_i} | Factor of sensibility A_i | Contribution to the standard uncertainty u_{x_i} ; uncertainty u_{CT_R} |
|--|---------------------|--------------------------------|-----------------------------|---|
| Slope b | -0.00618m | 0.00072m | 27083K/m | 0.00072K |
| Resolution of the spectrometer | - | $-1.9 \cdot 10^{-10}$ m | $3.199 \cdot 10^{10}$ K/m | $-1.9 \cdot 10^{-10}$ K |
| Resolution and experiences of the user | - | $5 \cdot 10^{-10}$ m | $3.199 \cdot 10^{10}$ K/m | $5 \cdot 10^{-10}$ K |
| T_R | 398.8K | - | - | 25.94K |

Table 1: The balance table of uncertainties for the rotational temperature determination from the pyrometric line. The slope is calculated for the significant level 0.05. According the row TR we can say that the real value of rot. temperature lies in the interval 372.86–424.74K with the probability 68 %.

This research has been supported by the Czech Ministry of Education grant No. 2004/0150.

References

- [1] Vocabulary and Symbols, Part I: Probability and General Statistical Terms; ISO 3534–1, 1993. ČSN ISO 3534–1, 1994).
- [2] Z. Sládek, F. Vdoleček: Technická měření; VUTIUM, Brno, 1992.
- [3] J. Anděl: Statistické metody; MatFyz Press, Praha, 1993.
- [4] B. Gross: Technika plazmatu; SNTL, Praha, 1967.
- [5] I. Kovacs: Rotational Structure in the Spectra of Diatomic Molecules; Akademiai Kiado, Budapest, 1969.
- [6] M. Meloun, J. Militký: Kompendium statistického zpracování dat. Academia, Praha, 2002.
- [7] Z. Karpíšek: Applied Statistics; Brno PC-DIR Real, Brno, 1999.
- [8] <http://www.milansorm.cz/simul>

Teaching Chemometry and Good Laboratory Practice at Palacký University in Olomouc

David Milde

*Department of Analytical Chemistry, Faculty of Science, Palacký University
Tr. Svobody 8, 771 46 Olomouc, Czech Republic
david.milde@upol.cz*

The application of modern instrumental techniques in many chemical laboratories loads students and chemists with plenty of the results that are to be responsibly evaluated. In the beginning of the 21st century, the university graduates do not manage with using mean and standard deviation but should be able to apply present-day techniques in some statistical software besides good knowledge from their specialization in chemistry.

Nowadays at Faculty of Science there is one semester course in Good Laboratory Practice (GLP) that students take in 3rd year of their bachelor studies. This course is focused on the foundations of GLP (sampling, techniques of calibration and its evaluation, certified reference materials, ...), control charts, evaluation of uncertainty, validation of analytical method and design of experiments (DOE). Chemometry is taught in 1st year of graduate studies in 2 semester courses that are a follow up to a short statistical course in lectures of Mathematics. The first semester of chemometry, which is compulsory for students, deals with statistical evaluation of univariate data (exploratory and confirmatory analysis, hypothesis testing, ANOVA) and linear regression and calibration. The second course, which is optional, covers with non-linear regression and foundations of multivariate statistical analysis (factor analysis, PCA, cluster analysis).

All given courses consist from lectures and seminars in computer classroom with access to same statistical software (nowadays QC-Expert and Statistica). Studying materials for lectures and seminars are prepared for students on web site (<http://aix.upol.cz/~milde> - only in Czech). The details about the content of courses and experiences with teaching and the evaluation of students will be presented.

Miminizing the Effects of Multicollinearity in the Polynomial Regression

Milan Meloun

*Department of Analytical Chemistry,
Faculty of Chemical Technology, Pardubice University,
532 10 Pardubice, Czech Republic
milan.meloun@upce.cz, <http://meloun.upce.cz>*

Multicollinearity (strong correlations among independent variables) is characteristic in the commonly used ordinary polynomial regression models. The difficulties could result in a complete misinterpretation of the data. A principal component polynomial regression (PCR) of the data transformed by power transformation was chosen as a suitable method capable to cope with the aforementioned problems. This paper emphasizes the importance of understanding the nature of near-singularities in the data which might cause problems with ordinary least squares regression, and describes the algorithm of one biased regression method. Several statistical criteria for the selection of suitable bias as the mean quadratic error of prediction MEP, the predicted coefficient of determination R_p^2 and the Akaike information criterion AIC on the problem of parameter estimation in a polynomial regression model should be considered together. The proposed algorithm in S-Plus of generalized principal component regression $GPCR$ is demonstrated on a problem in clinical biochemistry: for the age dependence of epitestosterone EpiTe a polynomial regression model was built and the question answered as to whether age-related changes in the concentration of this steroid in men and women are significant. Significant differences were found between men and women in the course of the age dependence of steroid. In women, a significant maximum was found around the 30th year followed by a rapid decline while the maximum in men was achieved almost 10 years earlier and changes were minor up to the 60th year. As concerns the method of data analysis, principal component regression is very useful tool for the investigation of curvilinear dependencies especially in polynomial regression models.

Keywords: Biased linear regression; Mean quadratic error of prediction; Multicollinearity; Principal component regression PCR.

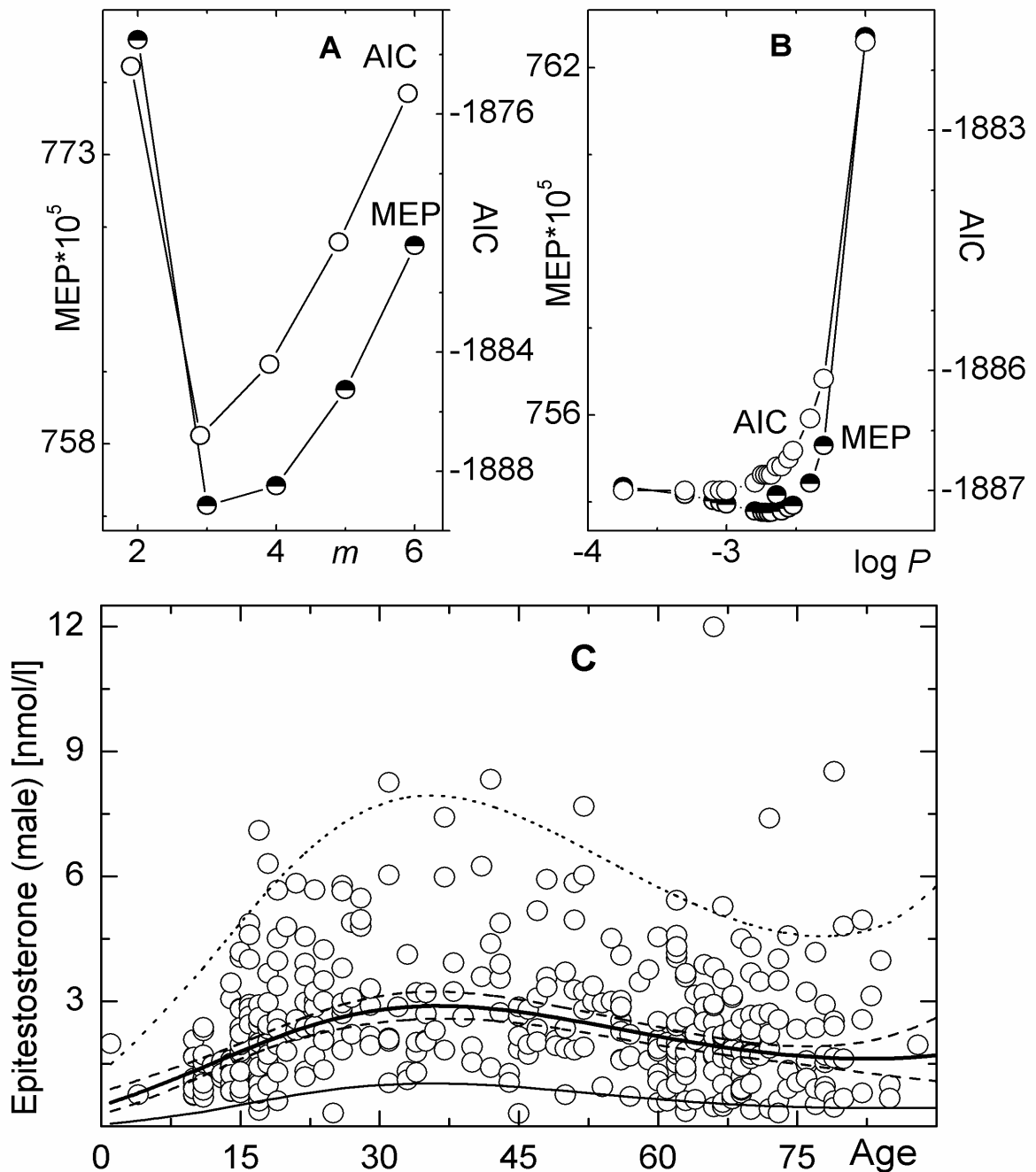


Fig. 1 Polynomial regression of the age dependencies of EpiTe in the serum of 211 females aged 10 - 77 years. Because of the skewed data distribution on the y-axis, the original data were transformed to the minimum skewness of the y values with power $\lambda = -0.13$. (A) The search for an optimal polynomial degree $m = 4$ led to one extreme when the ordinary least squares OLS and the mean error of prediction MEP on m or the Akaike information criterion AIC on m were used. (B) The search for the PCR optimal criterion value $P = 4.0 \times 10^{-6}$ separates statistically significant and insignificant parameter estimates when the method of generalized principal component and the mean error of prediction MEP on P or the Akaike information criterion AIC on P are used. (C) The curves of the mean prediction (the *solid line*) with the 95% confidence intervals for a prediction for the centre of gravity of the independent variable x_0 (the *dashed line closer to the mean prediction*) and the 95% Working-Hottelling confidence intervals of prediction for any x value (the *dotted line further from the mean prediction*) were obtained by the retransformation of the results to the original scale. All of the parameters of the polynomial were statistically significant (t -tests).

Consistency analysis for PLS with variable selection to diagnose a dehydration process (poster)

Manuel Zarzo, Alberto Ferrer

Department of Applied Statistics, Operations Research and Quality

*Polytechnic University of Valencia, Camino de Vera s/n; edificio I-3, 46022 ; Valencia (Spain),
mazarcas@eio.upv.es, aferrer@eio.upv.es*

In many cases, quality control of chemical batch processes is not easy, since a certain amount of the batches produced are out of specifications and the causes remain unknown. This problem occurs in an industrial chemical polymerisation batch process in 4 stages. The last one is the dehydration of the polymer. One of the quality parameters analysed in the final product is the residual water content, and in 15% of the batches it exceeds the upper tolerance limit. This process is controlled automatically with the information (temperatures, pressures, flows, etc.) acquired on line by means of 38 electronic sensors. The trajectories of these process variables have been provided for 68 batches, produced in two different periods. As the duration of the stages is not constant from batch to batch, alignment methods have been used in order to synchronise the trajectories, resulting a final matrix that contains thousands of aligned variables. The logarithm of the residual water content has been considered as response variable because this parameter can be modelled by a log-normal distribution. Fitting a PLS regression with this matrix, the first component is not statistically significant according to cross-validation. The same happens using only the 38 batches from the first period. But with the second data set (30 batches), a significant model is obtained ($Q^2=0.42$). To diagnose the causes of variability of the residual water content, the variable loadings in the first component from both PLS models have been plotted in the same chart, that has been examined to identify consistent runs (sequences of aligned variables with high loadings for both data sets). The highest loadings for the second data set correspond to the valve, temperature and pressure during the final dehydration, but a different pattern has been detected for the loadings from the first data set. These results show the importance of consistency analyses to diagnose chemical processes when batches belong to different periods, since the causes of correlation might not be the same for all batches.

Analysis of dynamic gas sensor response using chemometric methods (poster)

Manuel Zarzo¹, Abelardo Gutiérrez², Enrique Moltó²
mazarcas@eio.upv.es, gutierre@ivia.es, molto@ivia.es

¹*Department of Applied Statistics, Operations Research and Quality
Polytechnic University of Valencia
Camino de Vera s/n ; edificio I-3 ; 46022 ; Valencia (Spain)*

²*Instituto Valenciano de Investigaciones Agrarias (IVIA)
Carretera Moncada – Náquera Km. 4.5 ; 46113 Moncada, Valencia (Spain)*

Electronic noses are able to assess the quality of many foodstuffs and commodities if they are properly calibrated with statistical methods. An electronic nose has been used to assess 26 samples of olive oils prepared with increasing intensities of rancid and winey attributes, that have also been evaluated by a panel of trained tasters to quantify the intensity of the defect. For every sample, the device records during 8 minutes the evolution of an electric signal from 8 metal oxide semiconductor gas sensors, with a sampling rate of 1 Hz. Hence, data are structured in a 3-way matrix (26 samples x 8 sensors x 480 seconds). Due to the high number of variables, usually different parameters are extracted from the signals (extreme and relative values, slopes, etc.) and analysed with classical regression methods. Since variables are strongly correlated, multivariate methods based on projections to latent structures can be used. First, this data matrix has been analysed with Unfold Principal Component Analysis (U-PCA), centring the data and scaling to unit variance. The score plot clearly reflects that samples with high intensities of rancid defects are discriminated from the winey ones. Afterwards, fitting a PLS regression to the unfolded matrix (U-PLS), a predictive model is obtained that classifies properly the samples according to the type of the defect, except for low intensities. This is a rapid and simple method to analyse dynamic gas sensor response, that also allows the detection of outliers and the identification of those gas sensors with better discriminant capacity. The development of an efficient methodology to calibrate electronic noses, based on sample preparation, measurement procedure and a proper data analysis, is a crucial step to ensure a long-term operation in these devices.

Number of components using modified PCA scree plot in spectroscopy

Tomáš Syrový, Milan Meloun

Department of Analytical Chemistry, Faculty of Chemical Technology,

University of Pardubice,

Naměstí Čs. Legii 565, 532 10 Pardubice, Czech Republic

e-mail: tomas.syrovy@upce.cz, milan.meloun@upce.cz

The determination of the number of components in a mixture is an important tool for qualitative and quantitative analysis in spectroscopy. The accuracy of selected indices for an estimation of the number of components that contribute to a set of spectra was critically tested on experimental data sets of protonation equilibria of drugs using the INDICES algorithm in S-Plus. Methods are classified into precise methods and approximate methods. Methods are based on the first criterion concerning the procedure on finding the point where the slope of the indicator function $PC(k) = f(k)$ changes so called scree plot. Besides the first criterion applied, indicator function $PC(k)$ of precise methods are also based on a comparison of an actual index $PC(k)$ of method used with the experimental error of the instrument used, $s_{\text{inst}}(A)$. The precise methods are *Kankare's residual standard deviation*, $sk(A)$, *Residual standard deviation*, $RSD(k)$, *Average error criterion*, $AE(k)$, *Bartlett χ^2 criterion*, $\chi^2(k)$. Improved identification uses the second or third derivative function for some indices, namely when the number of component in the mixture is higher than three and when, due to large variations in the indicator values even at logarithmic scale, the indicator curve does not reach an obvious point where the slope changes. The *derivative criteria* $SD(k)$ are based on the point where the slope changes and reaches a maximum. The *third derivative* $TD(k)$ value crosses zero and reaches a negative minimum which can be used as a criterion. The change in slope can also be found by calculating the *derivatives ratio* $ROD(k)$. Ideally $ROD(k)$ should have a maximum at the point where $k = p$. A more difficult problem is to deduce the number of components without relying on an estimation of the instrumental error of absorbance, $s_{\text{inst}}(A)$; then the first criterion only remains. Empirical functions are *Exner function* $\psi(k)$, *Scree test*, $RPV(k)$, *Imbedded error function*, $IE(k)$, *Factor indicator function*, $IND(k)$, *Ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA*, $RESO(k)$.

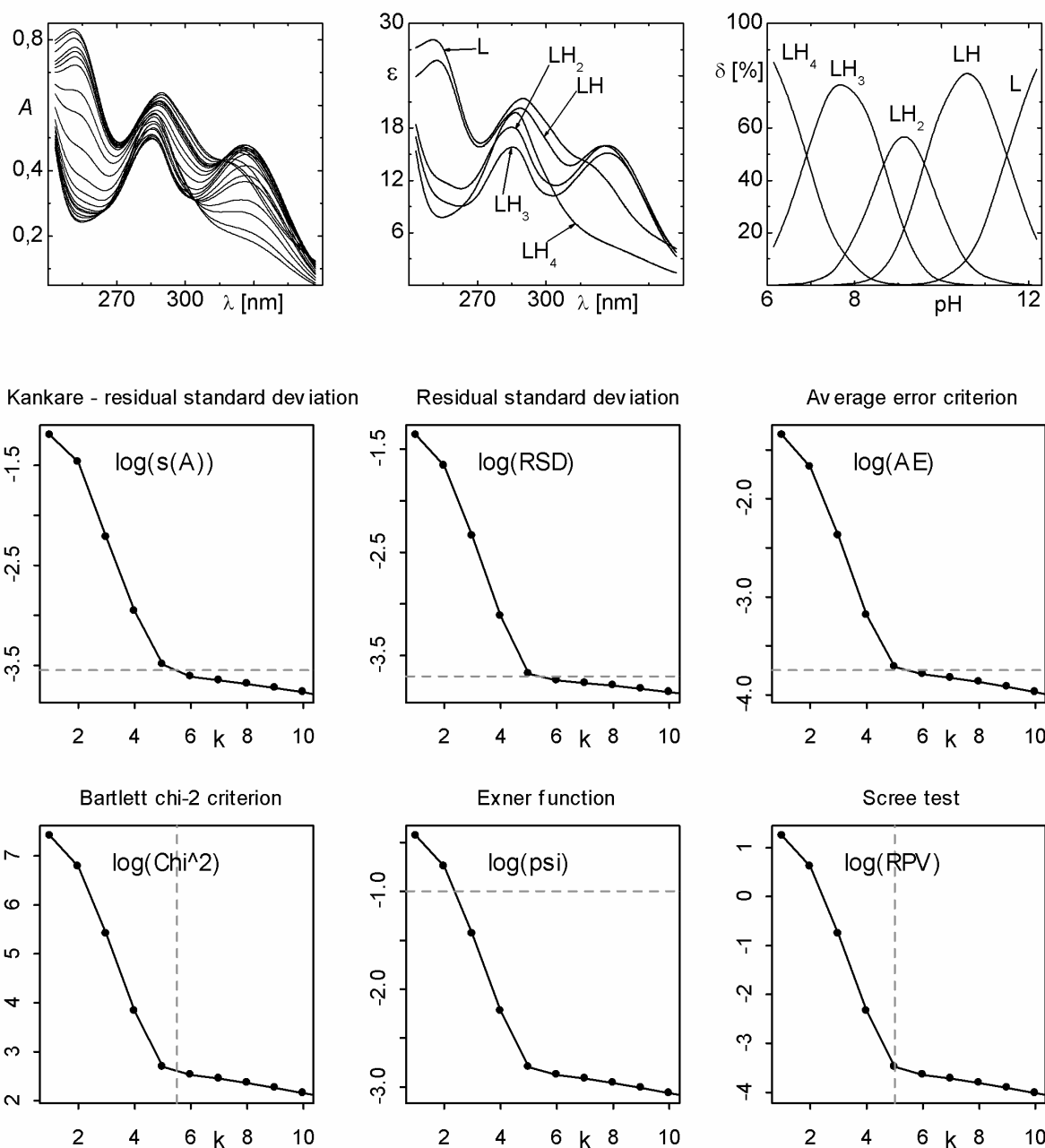


Fig. 1 Protonation equilibria of silybin is presented on pH-absorption spectra at 25EC; The spectra of molar absorptivities vs. wavelengths for all variously protonated species; Distribution diagram of the relative concentrations of all of the variously protonated species. The logarithm dependence of 6 indices methods as a function of the number of principal components k for the pH-absorbance matrix: 2nd row: Kankare's residual standard deviation $s_k(A)$, Residual standard deviation RSD , Average error criterion AE , 3rd row: Bartlett χ^2 criterion, Exner ψ function, Scree test RPV .

Keywords: Factor analysis; Number of species; Determining the number of components

[1] M. Meloun, J. Čapek, P. Mikšík, R.G. Brereton, *Critical comparison of methods predicting the number of components in spectroscopic data*, Anal. Chim. Acta 423 (2000) 51–68.

[2] M. Meloun, T. Syrový, A. Vrána, *Determination of the number of light-absorbing species in the protonation equilibria of selected drugs*, Anal. Chim. Acta, 489 (2003) 137 - 151.

Simultaneous Voltammetric Chemometric Determination of the anticancer drugs: Tarabine PFS, Adriblasinta and Methotrexate.

Deia Abd El-Hadi and Nagwa AboEl-Maali*

Department of Chemistry, Faculty of Science, Assiut University, 71516 Assiut, Egypt

Abstract

Chemometric approaches such as classical least squares (CLS), principle component regression (PCR), partial least squares (PLS) and iterative target transformation factor analysis (ITTTFA) have been applied to the simultaneous determination of mixtures of the anticancer drugs Tarabine PFS, Adriblastina and Methotrexate by Osteryoung Square Wave Voltammetry (OSWV) at the in-situ mercury film electrode. The conventional and first-derivative voltammograms of these mixtures were used to perform the optimization of the calibration procedure by chemometric models. At pH 7.4 (Phosphate buffer), the proposed method was applied satisfactorily to the determination of a set of synthetic mixtures of these anticancers. The obtained results of the different chemometric approaches are discussed and compared. Quality Control Charts are also constructed.

Exploratory data analysis - Addressing the context in which chemometrics works

Ayobami David Adegbola

University Of Ibadan, Nigeria, Department Of Chemistry, Dugbe Ibadan Nigeria

ayoadebola@yahoo.com

Keywords: exploratory data analysis, multivariate screening, induction, data mining, induction

Abstract

The instrumental revolution in chemistry and physics has produced a wide range of destructive and non-destructive identification and separation methods, which can be used as fast and inexpensive multivariate screening methods in considering laboratory as well as real life data. Within the limits of the analysis these methods may give a chemical-physical data fingerprint of a product or process which may contain more information than scientists as a collective may hypothesize.

The hypothesis-generating exploratory data analysis was developed in the social and economic sciences, starting with the principal component analysis (PCA) algorithm. Its application is discussed in examples from the food industry based on data from multivariate screening methods including multiblock/multiway extensions of PCA such as PARAFAC and Tucker which have more recently been applied in chemometrics.

Exploration can be pursued as early as in the formulation phase of a problem as a pre-project, dynamically employing a multivariate screening method (e.g. based on spectroscopy) covering the problem space with the generally held hypothesis that this strategy will produce a data base which could embrace the problem. The covariate latent structure of the resulting data set containing different blocks of data at different context levels of biological or technological organization is presented graphically as principal components (PCs) in a cognitively interpretable form. The identification and the naming of the PCs acts as a stimulus for the intuition of the investigator, occasionally generating a new fresh hypothesis which later could be tested by experimental design based on the factors revealed.

It is concluded that the exploratory inductive strategy using modern technology now available makes a new, largely independent channel of information accessible to classical, deductive, normative science, making possible a fruitful dialogue. The role of exploratory chemometrics at present and in the future with the aim to economize the input of resources in research is discussed using examples. There is a tendency that the prevailing normative strategy leads to specific problems getting solved, while multifactorial problems accumulate. A much higher success rate should be guaranteed by allowing the financing of exploratory pre-projects based on multivariate screening methods and exploratory data analysis as a complement. The challenge of how to launch such exploratory chemometrics is discussed.

Statistical Software System
SixSigma QA/QI ISO9000
Quality Control
Implementation
Consulting
Data Analysis
Solutions
Training
Software

TriloByte[®]
 Ltd.
 STATISTICAL
 SOFTWARE
 www.trilobyte.cz

**Your Data Deserve
 Statistical Analysis...**



- Elementary Statistics
- Probabilistic Models
- Analysis of Variance
- Correlation
- Transformation
- Error Propagation
- Uncertainty
- Simulation
- Samples Comparison
- Statistical Testing
- Optimisation
- Response Surfaces
- Calibration
- Shewhart Control Charts
- Modern Control Charts
- Capability
- Acceptance Sampling
- Linear Regression
- Linearized Regression
- Robust Regression
- Nonlinear Regression
- Multivariate Analysis
- Probability Calculator
- Graphical Visualization
- Documentation
- Administration of Analyses
- Data Management
- Help System
- User Manuals
- Seminars, Education
- Consulting Solutions

Connecting.
Analysing.
Modelling.
Explaining.
Understanding.
Informing.
Deciding.
Solving.
Improving.
Profiting.
...QC.Experting.

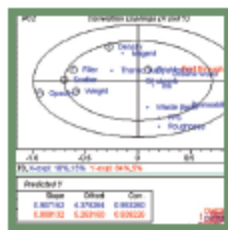
...QC-Expert.
The Statistical System.

www.trilobyte.cz

Leading software in Multivariate Data Analysis



The Unscrambler®



Pioneering Methods

The Unscrambler® is Continuously developed in close cooperation with the world leading chemometricians and industrial partners, to bring you the latest and most efficient advances in data analysis.

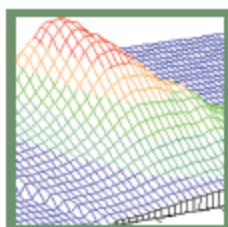
Data analysis tools for Research and Industry

The Unscrambler® offers powerful methods like PCA, PLS, Three-way PLS regression and experimental design for your application in product development, process control, quality control, research and all kinds of projects involving analysis of small or large amounts of data.



Analysis Methods

- Descriptive statistics (Mean, Standard Deviation, Box-Plot, Skewness, kurtosis...)
- Principal Component Analysis (PCA)
- Regression (MLR, PCR, PLS-R, 3-way PLS-R) and Prediction
- Classification (SIMCA, PLS-DA)
- ANOVA and Response Surface ANOVA
- Validation options: Leverage Correction, Cross-Validation (freely choose number of samples per segment), Test Set
- Variable Scaling options: scaling is free on each variable. Suggested options: auto-scaling, constant, passify
- Interaction and Square terms can be included in all models

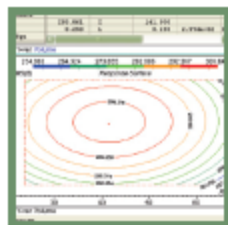


Smart tools for Analysis

- Automatic detection of significant X-variables in PCR and PLS-R
- Model Stability in PCA, PCR and PLS-R
- Automatic outlier detection in PCA, MLR, PCR, PLS-R and Prediction
- Interactive analysis:
 - ♦ Mark samples and/or variables on plots
 - ♦ Recalculate with or without Marked samples or variables
 - ♦ Recalculate with Passified Marked or Unmarked variables
 - ♦ Extract Data from Marked or Unmarked
- Automatic Pretreatments in Prediction and Classification
- Interactive help



Free trial versions of the Unscrambler's products family are available for download on www.camo.com



WEB Seminars:

Join us for a free, on-line, web seminar where you will discover features, benefits, and user friendliness of the Unscrambler. Our Schedule is available on <http://www.camo.com/rt/news/Webseminars/websemoslo>






Training program :

CAMO provides professional training on multivariate data analysis and experimental design across the United States and Europe. This includes hands on training on the The Unscrambler® product. Get more information on Camo's Web page.

Camo Process AS
Nedre Vollgate 8, N-0158 Oslo, Norway
Tel: +47 2239 6300 Fax: +47 2239 6322
E-mail: camo@camo.no Web: www.camo.no

Sponsors of the conference

We wish to thank sponsors of the conference:

| | | |
|---|--|---|
|  TriloByte Statistical Software |  Camo, The Unscrambler Software | |
|  University of Pardubice |  Technical University of Liberec |  Czech Chemical Society |



