The
North American Chapter
of the
International Chemometrics Society

Newsletter # 22 NAmICS January 2002

Barry M. Wise
President
Eigenvector Research, Inc.
830 Wapato Lake Road
Manson, Washington 98831
bmw@eigenvector.com

Patrick Wiegand
President-Elect
Union Carbide Corporation
P.O. Box 8361
South Charleston,
West Virginia 25303
wieganpm@ucarb.com

Margaret A. Nemeth
Secretary
Monsanto Company
800 N. Lindbergh Boulevard
St. Louis, Missouri 63167
MANeme@ccmail.Monsanto.com

Aaron J. Owens
Editor-in-Chief
DuPont Company
POB 80249
Wilmington, DE 19880-0249
aaron.j.owens@usa.dupont.com

Neal B. Gallagher
Treasurer
Eigenvector Research, Inc.
830 Wapato Lake Road
Manson, Washington 98831
nealg@eigenvector.com

Ronald E. Shaffer
WebMaster
GE Corporate R&D
Building K1, Room 3B12
P. O. Box 8
Schenectady, New York, 12301
shaffer@crd.ge.com

David Duewer
Membership Secretary
National Institute of Standards
and Technology
Gaithersburg, Maryland 20899
David.Duewer@NIST.gov

Table of Contents

Thanks to

MATFORSK

for supporting the duplication
and mailing of the newsletters.

MATFORSK has a long tradition in chemometrics, starting in the late 70's with multivariate calibration in NIR applications, and some years later multivariate analysis of sensory data was started up. Today, the group of chemometrics and statistics consists of about 10 persons divided among research scientists and Ph.D. students. More info can be found at

http://www.matforsk.no/web/statchem.nsf

Guest editorial

In a recent review in Trends in Biotechnology it is claimed: "Too often, data mining activities are simply large-scale applications of poorly understood methods to poorly understood data". Well, what is data mining anyway? A quick search on the Internet revealed that many companies within data mining rely on methods such as cluster analysis, factor analysis, neural networks, decision trees. Interestingly enough, regression is often not mentioned. It is because this term generates bad vibrations among the "data miners" and their customers?

For two years now, I have work with chemometrics at MATFORSK; Norwegian Food Research Institute. The applications span from bioinformatics to analysis of consumer related data in a variety of different projects, and it involves planning experiments, sampling from biological systems and analysis of considerable amounts of data, often in blocks, from measurements on raw materials, processes and products. The contributions in this issue cover some aspects regarding the role of chemometricians as "information extractors" in the era of vast amounts of data.

What is the role of the chemometrician today? – and what is a chemometrician? I am inclined to support the view that "chemometrics is what chemometricians do" (see also Jerry Workman's contribution in this issue). At MATFORSK we often work with sensory analysis in food research, and why should we refrain from analyzing accompanying data from questionnaires related to consumer's preferences, attitudes, eating habits etc.? As sound users of statistical methods, we should initiate proper use of them also in other fields of science. Now is the time to apply versatile methods that we have found to be apt for huge data sets in chemistry to other fields, like genetic data. I think that chemometricians can contribute greatly within data mining, being it bioinformatics or CRM. Many variables, relatively few objects, missing data, outliers and many responses are situations we are used to handle within chemometrics. The use of resampling/sub-sampling methods to make models on some representative sets of objects is another issue that fits into chemometric thinking.

All the best for the new chemometric year - and may your data be with you!

Frank Westad
MATFORSK
frank.westad@matforsk.no

Listserv changes

The University of Maryland is terminating the IBM VM/CMS computer known as umdd.umd.edu. This means, among other things, that Listserv lists will be moved to a new computer. The ICS-L Listserv list (subscriptions and list archives) has been moved. All postings should now be sent to:

ics-l@listserv.umd.edu
(NOT ics-l@umdd.umd.edu)

Listserv administrative mail (subscribe, unsub-scribe, changesubscription options, etc.) should be sent to:

listserv@listserv.umd.edu
(NOT listserv@umdd.umd.edu)

Note that mail from the list is now sent through LISTSERV.UMD.EDU, instead of UMDD.UMD. EDU. If you filter your messages based on the email address on the TO: line or REPLY-TO: line, you may need to modify your filters for this list.

The new server (operating on a Unix system) is the current version of the Listserv system, much newer than the version we were able to use on the UMDD system. One of the nice new features for us is a web interface to the Listserv server. The Listserv server "home page" may be accessed at:

http://www.listserv.umd.edu

The web interface for this list may be accessed at:
http://www.listserv.umd.edu/archives/ics-l.html

Chemometrics in a Network Economy

Jerry Workman, Jr.

Kimberly-Clark Corp., Analytical Science and Technology Group, West Research and Engineering, 2100 Winchester Road, Neenah, Wisconsin 54956 U.S.A.

This paper is an excerpt of a presentation made at the Fourth International Conference on Environmetrics and Chemometrics, Las Vegas USA September 2000. For a more complete discussion please refer to the paper, "The State of Multivariate Thinking for Scientists in Industry: 1980-2000," to be published in Chemometrics and Intelligent Laboratory Systems 2001.

1. Introduction

Chemometrics has enjoyed tremendous success in the areas related to calibration of spectrometers and spectroscopy-based measurements. Chemometric-based spectrometers have been widely applied for process monitoring and quality assurance. However, chemometrics has the potential to revolutionize the very intellectual roots of problem solving. Are there barriers to a more rapid proliferation of chemometric based thinking, particularly in industry? What are the potential effects of chemometrics technology and the New Network Economy (NNE), or simply the network economy, working in concert? Who will be the winners in the race for faster, better, cheaper systems and products? These questions are discussed briefly in terms of the principles of the network economy and in the promise of chemometrics for industry. What then is the state of chemometrics in modern industry? Several powerful principles are derived from an evaluation of the network economy and chemometrics which could allow chemometrics to proliferate much more rapidly as a key general problem solving tool.

In chemistry, one's ideas, however beautiful, logical, elegant, imaginative…are simply without value unless they are actually applicable to the one physical environment we have; in short, they are only good if they work. – R. B. Woodward

Relating the theme of chemometrics to a busy technical community

Twenty years after the term chemometrics was freshly "minted," by Bruce Kowalski and Svante Wold, the chemometrics community still seems to be searching for a universal definition and a clear identity. This paper begins by examining several definitions for chemometrics, the clarity of these definitions, and the message communicated to the industrial community.

"Chemometrics is what chemometricians do." Anon.

"Chemometrics has been defined as the application of mathematical and statistical methods to chemical measurements [1]."

"Chemometrics is the chemical discipline that uses mathematical and statistical methods for the obtention in the optimal way of relevant information on material systems [2]."

"Chemometrics developments and the accompanying realization of these developments as computer software provide the means to convert raw data into information, information into knowledge and finally knowledge into intelligence [3]."

"Analytical chemistry has been called a science without a theory. Some say that the theories and principles of analytical chemistry have been handed down from other branches of science. Developments in chemometrics that are beginning to effect instrument design and specify the limits of analysis are shaping the foundation for this science....research in chemometrics will contribute to the design of new types of instruments, generate optimal experiments that yield maximum information, and catalog and solve calibration and signal resolution problems. All this while quantitatively specifying the limitations of each instrument as well as the quality of the data it generates [4]."

"Chemometrics, the application of statistical and mathematical methods to chemistry...[5]"

"Chemometrics is the discipline concerned with the application of statistical and mathematical methods, as well as those methods based on mathematical logic, to chemistry [6-9]."

"Chemometrics can generally be described as the application of mathematical and statistical methods to (1) improve chemical measurement processes, and (2) extract more useful information from chemical and physical measurement data [10]."

"Chemometrics is an approach to analytical and measurement science based on the idea of indirect observation. Measurements related to the chemical composition of a substance are taken, and the value of a property of interest is inferred from them through some mathematical relation [11]."

"Chemometrics (this is an international definition) is the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments; and (b) to provide maximum chemical information by analyzing chemical data [12]."

From these definitions we are left with a few nearly irrefutable facts: chemometrics involves chemistry and math….most probably data, and possibly sensors and measurements of processes.

Whatever the clear and present definition of chemometrics is, the industrial understanding of it is that it is complicated, it requires computers and that it could possibly be beneficial. However we are not exactly certain how or why it would be an advantage to use chemometrics. The fact is that chemometrics allows us to take off the shelf data, which many institutions have been generating *ad infinitum*, and "wring it out" to remove all the information content. This information can further be scrutinized to obtain real knowledge of processes and measurements. Knowledge for optimized and new products, processes, intellectual property estates, at reduced cost.

Chemometrics is so often linked with Process Analytical Chemistry, again defined by Kowalski "as the discovery and development of new and sophisticated analytical methods for use in-line as an integral part of automated chemical processes [13]." Some have said that process analytical chemistry is 90% hardware and 10% chemometrics. To an engineer that quantitative statement means one may be able to do without it. What we have then is a process – we make measurements – we collect data – we use chemometrics to obtain information – we review the information and attain real knowledge. If chemometrics is difficult to clearly define and communicate, what are its advantages and disadvantages?

2. Advantages of chemometrics

What, then are the clear advantages of chemometrics?

(1) Chemometrics provides speed in obtaining real-time information from data;
(2) It allows high quality information to be extracted from less resolved data.
(3) It provides clear information resolution and discrimination power when applied to second, third, and possibly higher-order data.
(4) It provides methodology for cloning sensors – for making one sensor take data "precisely" as another sensor.
(5) It provides diagnostics for the integrity and probability that the information it derives is accurate.
(6) It promises to improve measurements.
(7) It improves knowledge of existing processes.
(8) It has very low capital requirements – it's cheap.

In summary, it provides the promise of faster, cheaper, better information with known integrity. In addition, it is common sense to know that math is cheaper than physics – that computer programs that can solve problems that traditionally have required extensive hardware developments and advances; it represents a superior approach. We see then, that intelligence can replace physical and material solutions, much as the digital chip replaces the mechanical clock works. This is an important theme to further develop.

Recent reviews describing the remarkable proliferation of near infrared, infrared, and Raman chemometrics-based analyzers for use in process analysis are given in reference [14]. The references found in this review and many others cite several thousand cases where chemometrics was applied to calibration of sensors to analyze complex chemical mixtures and used for on-line or at-line analysis. Without chemometrics most, if not all, of these applications would not have been

possible. In fact, multivariate calibration is commonly accepted for process, research, and quality use applications throughout the world of spectroscopy.

3. Disadvantages of chemometrics

The perceived disadvantage of chemometrics is that there is widespread ignorance about what it is and what it can realistically accomplish. The notion that 'many people talk about chemometrics, but there are relatively few actually using it for daily activities and major problem solving in industrial situations.' This science is considered too complex for the average technician and analyst. The mathematics can be misinterpreted as esoteric and not relevant. And most important for industry, there are a dismal lack of official practices and methods associated with chemometrics.

Chemometrics requires a change in one's approach to problem solving from univariate to multivariate thinking; since we live in an essentially multivariate context. From pondering over spreadsheets to actually analyzing the data for its full information content. The old scientific method is passing away; a new scientific method is arising from its ashes. A new method requiring not a thought ritual, but rather a method involving many inexpensive measurements, possibly a few simulations, and chemometric analysis. The new method looks at all the data from a multivariate approach, whereas the old method requires the scientist's assumed powers of observation, from a univariate standpoint, to be the key data processor.

The Old Scientific Method (used for hundreds of years)
(1) Stating the problem
(2) Forming the hypothesis
(3) Observing and Experimenting
(4) Interpreting Data (traditionally univariate – pondering stage)
(5) Drawing Conclusions

The New Scientific Method (for routine problem solving)
(1) Measure a process (any chemical phenomenon or process)
(2) Analyze the Data (multivariate analysis)
(3) Iterate if necessary
(4) Create and test model
(5) Develop fundamental multivariate understanding of the process

Industry relies on approved and accepted methods which can easily be defended in a court of law. The implementation of methods must involve a minimum of risk to the user and to the organization sponsoring the user. Historically, most NIR papers use Ordinary Least Squares to compare predicted NIR results against laboratory results. However, some regulatory groups have questioned the use of least squares and associated multivariate calibration for analytical methods involved with product release or compliance.

4. Lessons for chemometrics from the network economy

Rules of the NETWORK ECONOMY [15]

To start, let's look at two provocative statements relating to the network economy: "Give it away and it becomes priceless…keep it for yourself and it becomes worthless, " and "One fax machine is worthless, two are extremely valuable, many are priceless…." In the network economy increased complexity is the friend of confusion and chaos. The average person remembers 7 ± 2 objects per human byte and the modes of communication provide multi-channel competition for any concept. Among other means of communication, there is mobile connection commerce, Internet commerce, direct print, television, radio, telephone and fax commerce, and direct human contact. Concepts must then be clearly formulated and communicated to have any meaningful impact.

To be fast and first requires risk taking, risk by definition has a high failure rate. The lesson is to expend energy reducing the cost of risk, not the rate of risk. One must make it more expensive to be slow than wrong. So what are the laws of the network economy?

(1) The world is moving toward connectedness
(2) Services become more valuable the more plentiful they are
(3) Networked systems grow exponentially
(4) Success becomes infectious
(5) Value explodes with membership
(6) Cost goes down the better and more valuable the services are

(7) The more of something given free the more valuable it becomes - wealth feeds off ubiquity
(8) Allegiances move away from organizations and toward networks
(9) Devolution is essential – grassroots and bottoms-up (users) are in control
(10) A move from atoms to bits – smaller, smarter electronic systems over mechanical solutions
(11) Sustainable disequilibrium – constant change is the order of things
(12) Find the right task, not how to do the wrong task better.

Assumption to knowledge ratio causes more problems than ever before: (1) In the network economy, experience can be your worst enemy – it can lie to you about today's reality, (2) technologies must maximize learning rate while minimizing cost, (3) technological approaches must reduce the cost of failure, not the rate of failure. In summary, you can fail often if you fail cheaply. H*uman attention* is one of the key resource problems in the network economy, making clear communications one of the key aspects of successful attention getting.

5. Summary: Calibration is not all that chemometrics has to offer

Calibration of infrared and near infrared spectrometers has been far and away the most noted use of chemometrics in industry. But is this the best and most desirable use of this powerful technology? Do industrial managers get excited about calibrating a sensor using a new optimized technique, or in answering the question as to whether PLS is better than PCR in this or that case? I think not. In fact the sensors are all supposed to make some measurement using all those optics and electronics in the box - and that is that! Most chemometric books discuss mostly the aspects of calibration with a few miscellaneous applications. Examples [16-17] are adequate to demonstrate this principle.

Now industry may be missing something, and with all deference to the authors of these texts, but this material is still uncodified (esoteric) and undiffused (not distributed for general consumption) and it looks like through chemometrics one mostly has tools for calibration – some for quantitative analysis, some for qualitative. Analysis techniques using the various forms of multivariate analysis and pattern recognition techniques are certainly available. What else is there? How about codifying the concept of applying the principles of multivariate thinking to every possible problem that suffers from "univariatism." Chemometric tools would be most powerful if used for discovery of principles where multivariate data is available to study important variables in processes and product performance, where understanding cause and effect leads to product improvements.

In conclusion, eight basic, but powerful principles might be derived from the network economy that relate to chemometrics:
(1) The commodity most lacking in today's network economy is human attention
(2) Value = Utility + Ubiquity
(3) A clear powerful message with results receives attention and communicates utility
(4) The easiest way to get the world using chemometrics is to solve their most urgent problems using the most parsimonious solutions
(5) Give these solution for free over the internet
(6) Once the value is noticed the techniques will proliferate more rapidly
(7) Then make more advanced tools and instruction available for solving data problems through standard commercial solutions
(8) Make web-based data and enhanced algorithms available for everyone (i.e., network the global PC community into chemometric – and multivariate - thinking).

The chemometrician is thus encouraged to apply multivariate thinking as a new means to routine problem solving for calibration and discovery. By applying multivariate problem solving approaches to both analysis and discovery an improvement in both the depth of discovery and the speed of discovery is possible. Solve urgent problems first using the simplest approach. Offer solutions to those who will cooperate in a network of discovery to provide ubiquity and synergy. Develop special tools and approaches to discovery that will lead to faster and more insightful work.

6. References

1. B. R. Kowalski, Chemometrics, *Analytical Chemistry*, 52 (1980)112R-122R.
2. I. E. Frank and B. R. Kowalski, Chemometrics, *Analytical Chemistry*, 54 (1982) 232R-243R.
3. M.F. Delaney, Chemometrics, *Analytical Chemistry*, 56 (1984) 261R-277R.
4. L. S. Ramos, K. R. Beebe, W. P. Carey, E. Sanchez, B. C. Erickson, B. E. Wilson, L. E. Wangen, B. R. Kowalski, Chemometrics, *Analytical Chemistry*, 58 (1986) 294R-315R.
5. S. D. Brown, T. Q. Barker, R. J. Larivee, S. L. Monfre, and H. R. Wilk, Chemometrics, *Analytical Chemistry*, 60 (1988) 252R-273R.
6. S. D. Brown, Chemometrics, *Analytical Chemistry*, 62 (1990) 84R-101R.
7. S. D. Brown, R. S Bear, Jr., and T.B. Blank, Chemometrics, *Analytical Chemistry*, 64 (1992) 22R-49R.
8. S. D. Brown, T.B. Blank, S. T. Sum, and L. G. Weyer, Chemometrics, *Analytical Chemistry*, 66 (1994) 315R-359R.
9. S. D.Brown, S. T. Sum, and F. Despange, Chemometrics, *Analytical Chemistry*, 68 (1996) 21R-61R.
10. J. J. Workman, P. R. Mobley, B. R. Kowalski, R. Bro, Review of Chemometrics Applied to Spectroscopy: 1985-95, Part I, *Applied Spectroscopy Reviews*, 31 (1996) 73-124.
11. B. Lavine, Chemometrics, *Analytical Chemistry*, 70 (1998) 209R-228R.
12. B. R. Kowalski, in a formal CPAC presentation, Neenah, Wisconsin, December 12, 1997
13. B. R. Kowalski (personal communication) – Dec. 1997
14. J. J. Workman, Interpretive Spectroscopy for Near-Infrared, *Applied Spectroscopy Reviews* 31(1996) 251-320.
15. R. G. McGrath, Lecture notes and discussion from LECO program, Columbia University School of Business, July-Aug., 2000.
16. E. R. Malinowski, D.G. Howery, Factor Analysis in Chemistry, John Wiley & Sons, Inc., New York, 1980.
17. K. R. Beebe, R. J. Pell, M. B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, Inc., New York, 1998.

Genetic Algorithms in Feature Selection

Riccardo Leardi

Department of Pharmaceutical and Food Chemistry and Technology. University of Genova, Italy

1. Introduction

Since their presentation by Holland in 1975, the Genetic Algorithms (GA) have attracted a lot of curiosity. The goal (trying to simulate the evolutionary process of a living species) and jargon (using typical biological terms such as "gene", "chromosome", "mutation" and "cross-over" in the description of an algorithm) of GAs have helped to create something like an aura of mystery around them.

In that period, the main limitation to the real development of GAs in terms of applicability was the fact that the huge amounts of computation required by them could not be handled satisfactorily by the computers then available. For almost 20 years this has been the main problem for those who would have liked to apply them to their problems but did not have the possibility of accessing a suitable computer: for "common size" problems a mainframe would have been required, while for complex problems the computation time would have been too long even with the most powerful computers.

Since the beginning of the 1990s this major problem has been progressively removed, and nowadays every personal computer can be used to apply GA to easy/moderate-scale problems, while the mainframes allow one to tackle very complex problems such as those typical of molecular modelling. This is the reason why, after a first period in which the interest of the scientific community was focused mainly on the theory itself, the number of papers reporting applications of GAs to real problems and the number of scientists and of disciplines using them have been growing exponentially.

In 1993, the Journal "Science" [1] published a paper that gave a general presentation of genetic algorithms, some mathematical analysis about how

GAs work and how best to use them, and some applications in modelling several natural evolutionary systems, including immune systems. In 1995 an article in "Nature" [2] described a problem of molecular dynamics that had been successfully solved by a GA where conventional techniques had failed.

Several tutorials about GAs have been published in journals devoted to different research fields. As examples we cite those by Lucasius and Kateman [3-6], Hibbert [7], Shaffer and Small [8], Wehrens and Buydens [9] and Luke [10]. The set-up of the structure of a GA is a very critical point, and a guide leading to a good architecture is highly beneficial. Wehrens et al. [11-12] proposed a set of quality criteria to evaluate the performance of a GA, considering not only the best solutions suggested by the algorithm, but also the repeatability of the optimization and the coverage of the search space.

GAs have found widespread application in several fields involving regression problems. One of the most important steps in a calibration is the selection of the relevant variables. The size of the search domain (with v variables, 2^v-1 combinations are possible) and the presence of many local optima make GA one of the suggested methods. It is interesting to notice that several authors have published papers about feature selection by GAs, each of them using a different GA structure, sometimes rather far from the "standard" algorithm. This demonstrates the need to modify the algorithm according to the peculiarities of the problem to be solved. In the case of feature selection a chromosome is made by a very high number of genes (as many as the variables), each of them being just 1 bit long (0 = variable absent, 1 = variable present). Leardi et al. [13] use a simulated data set to show that a GA can always find the global maximum of a simple problem, in a time much shorter than the time required by a full search. Lucasius et al. [14] showed that a GA generally performs better than simulated annealing and stepwise regression; on the other hand, Hörchner et al. [15] demonstrated that simulated annealing can give the same results. Wise et al. also developed their GA for feature selection [16-17].

Wallet et al. [18] solve the problem of selecting a minimal model which correctly predicts the response by applying a GA using a two-criteria population management scheme.

Broadhurst et al. [19] applied GAs to pyrolisis mass spectrometry, with the goal of determining the optimal subset of variables to give the best possible prediction or determining the optimal subset of variables to produce a model with a predictive ability higher than or equal to a given value (both in MLR and in PLS models).

The method proposed by Bangalore et al. [20] leads to the selection of wavelengths and to the definition of the PLS model size. To do that, a "model size" gene taking on the integer value corresponding to the number of latent variables to be used in building the calibration model is added to the genes coding the presence or absence of each variable in the model.

Jouan-Rimbaud et al. [21] successfully applied a GA to a problem of wavelength selection for MLR calibration, while Arcos et al. [22] obtained a set of wavelengths able to perform a PLS calibration of mixtures of indomethacin and acemethacin, in spite of the fact that the two compounds have almost identical spectra.

A more complex optimization was performed by Shaffer and Small [23]. They apply a GA to the NIR analysis of glucose in biological matrices, optimizing at the same time five important variables: the position and width of the bandpass filter, the starting and ending points of the spectral range submitted to the PLS regression, and the number of latent variables employed in the calibration model.

In spectroscopic infrared imaging applied to discriminate between different materials, the selection of a limited number of spectroscopic wavelengths guaranteeing the optimal discrimination makes the acquisition and processing time much faster. This goal has been achieved by using GA by Van den Broek et al. [24]. Depczynski et al. [25] devised a method for multicomponent analysis by near infrared spectrometry by combining wavelet coefficient regression with a GA.

One of the main problems when applying feature selection to spectroscopical data is that the solution is often given by wavelengths scattered throughout the spectrum, instead of spectral regions as the solution given by spectroscopist. This problem has been tackled by Leardi [26], who modified its previous algorithm in order to force it

as much as possible toward the selection of contiguous wavelengths, in such a way that the final model can be more easily accepted also by spectroscopists. Although spectral data sets are the most common field of application of GAs for feature selection, owing to the very large search domain, also in the case of non-spectral variables some good results can be obtained, as reported by Aishima et al. [27].

Overfitting is the greatest risk when applying GAs to feature selection. This aspect has been taken into account especially by Leardi [28-29], Leardi and Lupiáñez González [30] and Jouan-Rimbaud et al. [31].

The ever-increasing number of papers in which GAs are applied to very different fields of chemistry and chemometrics shows the effectiveness and validity of this technique. The advantage of GAs over the "classical" techniques becomes greater the greater the complexity of the problem, especially owing to the good balance between exploration and exploitation. The results obtained by a GA can be highly improved after "hybridization" with a standard technique, typically having a very poor exploration ability and a very high exploitation potential. This allows one to define with great precision the global optimum, whose location has been effectively found by the GA. It is anyway to be highlighted that it is not possible to define a "best" GA architecture to be used in all applications, since the optimal structure is extremely problem-dependent.

3. References

1. Forrest S. 'Genetic Algorithms: principles of natural selection applied to computation'. *Science*, 1993, **261**: 872-878.
2. Maddox J. 'Genetics helping molecular dynamics'. *Nature*, 1995, **376**: 209.
3. Lucasius CB, Kateman G. 'Understanding and using genetic algorithms. Part 1. Concepts, properties and context'. *Chemom. Intell. Lab. Syst.*, 1993, **19**: 1-33.
4. Lucasius CB, Kateman G. 'Understanding and using genetic algorithms. Part 2. Representation, configuration and hybridization'. *Chemom. Intell. Lab. Syst.*, 1994, **25**: 99-145.
5. Lucasius CB, Kateman G. 'Gates towards evolutionary large-scale optimization: a software-oriented approach to Genetic Algorithms – I. General Perspective'. *Computers Chem.*, 1994, **18, 2**: 127-136.
6. Lucasius CB, Kateman G. 'Gates towards evolutionary large-scale optimization: a software-oriented approach to Genetic Algorithms – II. Toolbox description'. *Computers Chem.*, 1994, **18, 2:** 137-156.
7. Hibbert DB. 'Genetic algorithms in chemistry'. *Chemom. Intell. Lab. Syst.*, 1993, **19**: 277-293.
8. Shaffer RE, Small GW. 'Learning optimization from nature: Simulated Annealing and Genetic Algorithms'. *Anal. Chem.*, 1997, **69**: 236A-242A.
9. Wehrens R, Buydens LMC. 'Evolutionary optimization: a tutorial'. *TrAC*, 1998, **17, 4**: 193-203.
10. Luke BT. 'An overview of genetic methods'. *Genetic Algorithms in Molecular Modeling*, J. Devillers, Ed., Academic Press, 1996, 35-66.
11. Wehrens R, Pretsch E, Buydens LMC. 'Quality Criteria of Genetic Algorithms for structure optimization'. *J. Chem. Inf. Comput. Sci.*, 1998, **38**: 151-157.
12. Wehrens R, Pretsch E, Buydens LMC. 'The quality of optimization by genetic algorithm'. *Anal. Chim. Acta*, 1999, **388**: 265-271.
13. Leardi R, Boggia R, Terrile M. "Genetic Algorithms as a strategy for feature selection". *J. Chemom.*, 1992, **6**: 267-281.
14. Lucasius CB, Beckers MLM, Kateman G. 'Genetic algorithms in wavelengths selection: a comparative study'. *Anal. Chim. Acta*, 1994, **286**: 135-153.
15. Hörchner U, Kalivas JH. 'Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelengths selection'. *Anal. Chim. Acta*, 1995, **311**: 1-13.
16. Wise BM, Gallagher NB, Eschbach PA, Sharpe SW, Griffin JW. 'Optimization of prediction error using Genetic Algorithms and Continuum Regression: determination of the reactivity of automobile emissions from FTIR spectra'. *Fourth Scandinavian Symposium on Chemometrics (SSC4)*, Lund, Sweden, June 1995.
17. Wise MB, Gallagher NB, Eschbach PA. 'Application of a Genetic Algorithm to variable selection for PLS models'. *EUCHEM-Conference*, Gothenburg, Sweden, June 3-6, 1996.
18. Wallet BC, Marchette DJ, Solka JL, Wegman EJ. 'A Genetic Algorithm for best subset selection in linear regression'. *Proceedings of the 28th Symposium on the Interface*, 1996.

19. Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. 'Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolisis mass spectrometry'. *Anal. Chim. Acta*, 1997, **348**: 71-86.
20. Bangalore AS, Shaffer RE, Small GW. 'Genetic Algorithm-based method for selecting wavelengths and model size for use with Partial Least- Squares Regression: application to Near-Infrared spectroscopy'. *Anal. Chem.*, 1996, **68**: 4200-4212.
21. Jouan-Rimbaud D, Massart DL, Leardi R, De Noord O. "Genetic Algorithms as a tool for wavelength selection in multivariate calibration". *Anal. Chem.*, 1995, **67**: 4295-4301.
22. Arcos MJ, Ortiz MC, Villahoz B, Sarabia LA. 'Genetic-algorithm-based wavelengths selection in multicomponent spectrometric determinations by PLS: application on indomethacin and acemethacin mixture'. *Anal. Chim. Acta*, 1997, **339**: 63-77.
23. Shaffer RE, Small GW. 'Genetic Algorithm-based protocol for coupling digital filtering and Partial Least-Squares Regression: application to the Near-Infrared analysis of glucose in biological matrices'. *Anal. Chem.*, 1996, **68**: 2663-1675.
24. van den Broeck WHAM, Wienke D, Melssen WJ, Buydens LMC, 'Optimal wavelength range selection by a Genetic Algorithm for discrimination purposes in spectroscopic infrared imaging'. *Applied Spectroscopy*, 1997, **51, 8**: 1210-1217.
25. Depczynski U, Jetter K, Molt K, Niemöller A. 'Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm'. *Chemom. Intell. Lab. Syst.*, 1999, **47**: 179-187.
26. Leardi R. "Application of genetic algorithm-PLS for feature selection in spectral data sets", *Journal of Chemometrics*, 2000, **14**: 643-655.
27. Aishima T, Togari N, Leardi R. "Selection of aroma components to predict sensory quality of Kenyan black teas using genetic algorithms for multiple regression models". *Food Sci. Technol., Int.*, 1996, **2**: 124-126.
28. Leardi R. "Application of a Genetic Algorithm to feature selection under full validation conditions and to outlier detection". *J. Chemom.*, 1994, **8**: 65-79.
29. Leardi R. "Genetic Algorithms in feature selection", in '*Genetic Algorithms in Molecular Modeling*'. J. Devillers, Ed., Academic Press, 1996; 67-86.
30. Leardi R, Lupiáñez González A. "Genetic Algorithms applied to feature selection in PLS regression: how and when to use them". *Chemom. Intell. Lab. Syst.*, 1998, **41**: 195-207.
31. Jouan-Rimbaud D, Massart DL, de Noord OE. 'Random correlation in variable selection for multivariate calibration with a genetic algorithm'. *Chemom. Int. Lab. Syst.*, 1996, **35**: 213-220.

Parsimonious models in chemometrics

Frank Westad

MATFORSK
Norwegian Food Research Institute
www.matforsk.no

1. Introduction

The need for effective and safe methods for data analysis is increasingly important. Instrumentation and data collecting system generate huge amount of data. Examples are 2D or 3D analytical instruments, high-throughput screening, microarray, QSAR and imaging.

The parsimony principle [1] goes back several centuries. In a scientific context, a model with few parameters is to be preferred to other possible models if their interpretations and predictive abilities are similar. Various approaches exist for finding parsimonious models. A typical situation is some kind of best subset approach, which is discussed below in more detail. Another approach is to compress the original data, and estimate models on coefficients from the compressed representation of the data with methods such as splines and wavelets. This is useful both for storage and analysis of the data. Yet another approach to parsimony is to work with multiway, multiblock or multi-domain methods [2][3]. In this text we will not pursue this further, but focus on the task of finding relevant variables regardless if they are original or derived variables. It is also relevant to mention the newly developed o-PLSR [4] in this context. Another important issue that will not be dealt

with in this text is preprocessing of data, e.g. some kind of signal correction.

2. Some views on multivariate modelling

A vital aspect in all kinds of models is to estimate uncertainties for the model parameters. Since the truth is seldom known in empirical modelling, the uncertainties have to be estimated from a *sample* of the system or process under observation. Suffice for now the discussion about what is a training/calibration set and (independent) test set in terms of validation, let us assume that the objective is to make a model from the available data. Then, later, we might want to see if the observed system/process has changed over time or is behaving differently at another location and with other equipment. I think it is fair to say that model validation in one way or another is one of the important contributions from chemometricians in multivariate modelling. This should not be interpreted as we claim chemometrics "invented" model validation such as cross validation (CV), but it is an inherent part of most chemometric software packages.

Uncertainties may be estimated from resampling methods such as jackknifing (JK) and bootstrapping (BS). Jackknifing is closely connected to CV, the difference lies in whether the models with all objects or the mean of all individual models from the resampling should be regarded as the "reference". Since we are applying resampling for specific practical purposes such as variable selection, we feel it is more natural to look upon the model on all objects as the "reference". According to studies by Efron [5], the difference between these two is of order $1/N^2$, N being the number of objects.

Bootstrapping can be performed in various ways. The original approach was to resample from a population sample numerous times with replacement. From the theory, this will include 63% of the objects in one bootstrap sample. Another approach is to make a preliminary model, e.g. $\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$, and then draw individual pairs $(\mathbf{x},\varepsilon)$ randomly and generate new objects. Both these procedures are typically repeated 500 times. The results below are based on BS on original data.

Cross validation is intuitive in terms of stability of a model, formulated as the question "How is the model changing compared to the full model when some of the objects are taken out?" If then the model changes considerably, this deserves more detailed investigation as to *why* this occurred. This question is only common sense, but to quote Efron [5]: "*Good* simple ideas, of which the jackknife (JK) is a prime example, are our most precious intellectual commodity, so there is no need to apologize for the easy mathematical level".

How good are our estimates? Results from situations where the "truth" in terms of ANOVA is available show that both BS and JK estimates are close to ANOVA. Since JK has an aspect of validation, these estimates tend to be more conservative. An example is shown in Table 1 for the data set "Helicopter" taken from the literature [5]. The data are from a central composite design with four design variables and a blocking variable.

Table 1. Results for the Helicopter data

Variable	ANOVA	JK PC1	BS PC1
Block	0.613	0.671	0.583
Wing Area	0.960	0.970	0.962
Wing ratio	0.005	0.024	0.004
Body width	0.880	0.913	0.888
Body length	0.001	0.008	0.001

Variable selection in PCA

Variable selection is often thought of in a regression context, where one or more response variables (Y) are to be modeled by a set of predictor variables (X). However, the importance of finding the relevant variables is also vital in exploratory data analysis tools such as PCA, and for cluster analysis for that matter. In a recent review article about biotechnology it is claimed that "finding the few genes that are most responsible for the observed patterns in the data is a well-studied, but still unsolved, statistical problem".

We believe that jackknifing in PCA has a potential to solve some aspects of this "problem". It has now been applied on many different types of data. Below is an example from microbiology, where 26 bacteria have been tested to determine if they have the ability to ferment 28 different sugars, represented as binary data. PCA was employed, three PCs were found to be relevant. Significant variables on either or both PC 1 & 2 are marked as stars in Figure 1.

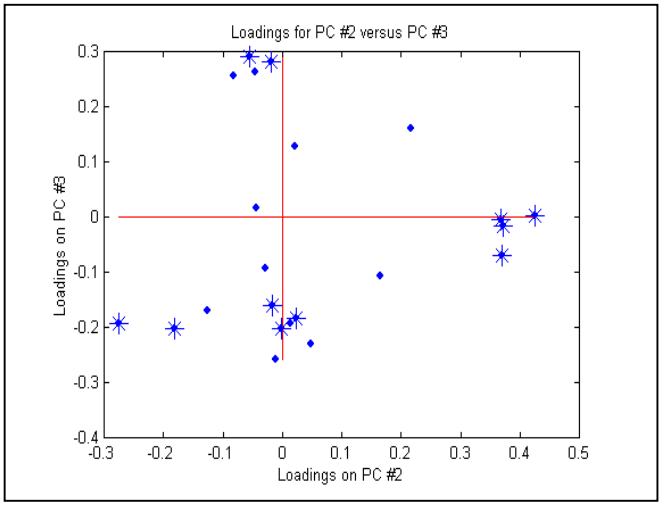

Figure 1 Loadings with significant variables marked

Although some of the variables have high negative loadings on PC2, they were found not to be significant. The reason is that these variables have the ability to ferment all sugars except one. The JK estimate thus warns us that the correlation between these variables and this PC is due to one object only; this is a way to assess the stability of the model. For data arising from experimental designs, the results so far show that significance variables found from JK estimates correspond well with Analysis of Effects from ANOVA.

Variable selection in regression

The majority of applications in variable selection concerns regression. Some of the most common methods for variable selection are:

- Stepwise approaches
- Genetic algorithms
- Methods based on uncertainty estimates (e.g. JK or BS).

There is a distinction between finding a model with good fit for a few variables, e.g. 3-7, and a model with all non-relevant variables removed. The JK and BS based approaches allow for removing all the non-relevant variables, but you are not restricted from taking out predictor variables because some of them are highly correlated. This of course is tied to the fact that most models do not have the best predictive nor interpretation abilities at full numerical rank. While we often see that JK uncertainty estimates from PLSR are higher than OLS estimates for full rank models, this is not contradictory to the general principle that the coefficients themselves are smaller for PLSR ("PLSR shrinks"). When the validation indicates that the optimal number of components is << numerical rank, then we *deliberately* want the estimates to be conservative for the OLS solution when employing e.g. cross validation.

Genetic algorithms have been shown to yield good results in a lot of different applications (see R. Leardi's contribution), but they require tuning of a number of parameters and they still are quite computer intensive for data with thousands of variables. An initial step where the non-relevant variables are removed by t-tests based on JK estimates of uncertainties will reduce the complexity.

The stepwise methods have a general problem in that relevant variables are never to be included because similar information is described by other variables, especially when the selection of variables is based on full rank models without proper validation. It is worth mentioning that this situation arises even when there are no numerical problems of matrix inversion due to collinearity. This is not to say that stepwise approaches do not give good models or predictions, for a small number of variables, typically 3-8, but one should monitor the performance with respect to the model rank.

The term "rank" with respect to a multivariate model deserves some comments, as "rank" has various facets:

1. The *numerical* rank. This rank is the one based on numerical computations, e.g. the number of components that can be computed without singularity problems.
2. The *statistical* rank. The important issue is here to find the optimal rank from a statistical criterion, preferably based on some proper validation method.
3. The *application specific* rank. This judgement is typically a combination of background knowledge, prediction ability, model complexity, and interpretation aspects. In most situations, this rank is lower than the statistical rank, i. e. the data-analyst tends to be more conservative.

Comparison of thousands, not to say millions of models to find THE best one may sound intriguing, but some questions arise in this context:

1. What is the stop criterion of not to include/exclude variables and/or components?

2. How do we judge if one model is better than another?

Variable selection by employing individual t-tests based on uncertainties estimated by jackknifing/CV and bootstrapping is not error-free. There is a danger to keep variables although they should have been removed or to overlook variables. Repeated random cross validation, e.g. 100 times, might be useful to see how many times a variable is included in the model. Another alternative is cross model validation [7][8]. Also, repeating the procedure iteratively helps in removing the non-relevant variables, as shown in [9]. This will in most cases also make it easier to assess the correct rank of the model.

Should JK or BS be employed? As the results in Table 2 indicate, the JK seems to be a more conservative method, and it has an inherent model validation. A study on the beer data reported in [9], showed that the BS estimates were smaller than JK, but thereby found a higher number of significant variables in the noise part of the spectrum. Table 2 shows the results when predicting 20 test samples. The results from calibration (not shown) were almost identical, but the JK approach gives significantly better prediction of the test samples.

Table 2: Results from modelling of beer data

Iter	#var/PC JK	RMSEP JK	#var/PC BS	RMSEP BS
0	0	2.44	0	2.44
1	926/6	0.74	926/5	0.72
2	162/6	0.20	375/6	0.58
3	90/9	0.20	300/8	0.44
4	16/7	0.20	237/8	0.46

It is a general experience that when a best subset model of e.g. five variables is the objective, many models will have more or less the same predictive ability.

In our pursuit of finding the best prediction (or classification) model based on variable selection, the ability to interpret the underlying structure might be distorted in the final model. It might seem illogical to discard variables we carefully have been observing/collecting. One alternative is to have two models:

1. One model with all variables where new samples are projected in the score plot, leverage and residuals can be interpreted etc.
2. One model for optimal classification/ prediction ability.

Instead of removing the non-significant variables, it is possible to down-weigh them with a factor of e.g. 10^{-4} so that their (numerical) influence in the model becomes small. Then, visualize them in the correlation loadings plot [10], which shows the correlation between the original variables and the components. This measure is invariant to how the variables were weighted during modelling. We end up with a model that has good predictive ability and at the same time all the variables can be interpreted.

3. Conclusions

The principle of parsimony has shown to be a fruitful approach in terms of variable selection to remove non-relevant variables. Jack-knifing is a conservative and less computer-intensive alternative to bootstrapping for estimates of parameter uncertainties in multivariate models. The list of variables sorted by significance is a good starting point for Genetic Algorithms and stepwise methods. Whenever variable selection in some form is employed, model validation is essential.

4. References

1. Seasholtz, M-B, and Kowalski, B. *Anal. Chim. Acta*, **277**, 165-177 (1993).
2. Reberg, J. O. and Martens, H. Patent W09629679 (1996).
3. Westad, F. and Martens, H. *Chemometrics and Intel. Laboratory Systems*, **45**, 361-370 (1997).
4. Trygg, J. and Wold, S. *Journal of Chemometrics* (in press).
5. Efron, B. *Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania*, ISBN 0-89871-179-7 (1982).
6. Box, G.E.P. and Liu, P.Y.T. *Journal of Quality Technology*, Vol 31, **1** (1999).
7. Anderssen, E and Martens, H. In prep. for *Journal of Chemometrics*.
8. Nørgaard, L. and Bro, R. Proceedings, Les Methode PLS. Symposium International PLS'99 187-202. (1999).
9. Westad F. and Martens. H. *Journal of Near Infrared Spectroscopy* **8**, 117-124 (2000).
10. Martens, H. and Martens, M. *Food Quality and Preference* **11**, 5-16 (2000).

Objective Data Alignment followed by Chemometric Analysis of Two-Dimensional Separations Initially with Retention Time Shifting on Both Dimensions

Carlos G. Fraga, Bryan J. Prazen and Robert E. Synovec

Center for Process Analytical Chemistry, Department of Chemistry, Box 351700, University of Washington, Seattle WA 98195, USA

1. Abstract

A retention time alignment preprocessing algorithm is presented that objectively corrects for run-to-run retention time variation on both separation dimensions of comprehensive two-dimensional (2-D) separations prior to application of chemometric data analysis algorithms. The alignment algorithm is easy to apply and robust. Thus, data from 2-D separation techniques such as comprehensive 2-D gas chromatography (GC x GC), liquid chromatography/liquid chromatography (LC x LC) and liquid chromatography/capillary electropho-resis (LC x CE) can be readily analyzed by various chemometric methods to increase chemical analysis capabilities. Complex samples can be more effectively studied.

2. Introduction

Comprehensive 2-D separations are ideally suited for the analysis of complex samples, and are emerging as powerful tools for chemical analysis [1-11]. Even with a large peak capacity, the probability of peak overlap in 2-D separations can be quite severe, especially for highly complex samples. Peak overlap becomes even more likely if one desires to speed up the analysis by designing a given separation method to provide a reduction in the run time. Thus, traditional methods of chromatographic and electrophoretic data analysis, such as peak height and peak area measurements, become less effective as the analyst moves into the realm of high-speed chemical analysis. The limitations brought upon by the likelihood of peak overlap can be overcome, to a large extent, by the implementation of appropriate chemometric methods. Essentially, chemometric methods will effectively enhance the resolving power of 2-D separation methods.

The Generalized Rank Annihilation Method (GRAM) is a chemometric method that resolves and quantifies overlapped peaks in common between sample and standard runs. Several papers cover the development of the GRAM algorithm in detail [12-14]. The specific requirements for GRAM analysis of comprehensive 2-D separations have been reported [1]. In order to use GRAM, the data comprising the peaks due to analytes and interferences present in the sample and standard must be bilinear or approximately bilinear. A bilinear peak is a 2-D peak that can be mathematically represented by the vector product of its elution profiles along each column separation axis, producing a data matrix. Generally, a bilinear peak in a contour plot appears as an elliptical (or circular) zone that has its major and minor axes aligned with the two separation axes. Contour plots from published GC x GC, LC x LC, and LC x CE papers depict 2-D peaks that appear bilinear [1-4, 8-11]. Using sample and standard matrices of data for sections of the 2-D separations that contain the analyte(s) of interest, GRAM calculates the pure elution profiles of overlapped 2-D peaks. Each 2-D peak can then be individually reconstructed using its respective two separation elution profiles. In addition, GRAM provides the concentrations for analytes in the sample relative to in the standard. The standard can be simply prepared using the original sample by the standard addition method [2].

Not all comprehensive 2-D separation techniques provide bilinear data. In particular, temperature programming simultaneously on both columns in GC x GC will not produce bilinear data. For this reason we have been developing a high-speed valve-based comprehensive GC x GC that has independently controlled temperature programming of both columns so the first column can be temperature programmed while the second column can be held isothermal at a desired temperature.

In order to use bilinear chemometric methods such as GRAM, the unwanted shifting of a 2-D peak's retention time(s) or migration time(s) between sample and standard runs must be objectively

corrected. Indeed, run-to-run retention time shifting has been a severe impediment to the use of chemometric methods on data collected from separation techniques. We have addressed the retention time alignment problem, and have previously developed an objective rank-based alignment method to correct retention time shifts along one time axis [15]. Rank alignment has been critically developed and successfully applied to GC x GC separations that need peak alignment on the first column time axis only [2-4]. Recently, we have now modified our alignment method to determine and correct the run-to-run peak shifting along both separation time axes. Here we report our initial findings. Correction of retention time shifting on both separation dimensions broadens the scope considerably for the use of chemometric methods, since most 2-D separation techniques produce run-to-run shifting on both dimensions that is significant and detrimental to chemometric applications, if not corrected.

3. Experimental

The original alignment method applies an iterative routine to determine the peak shift (along one time axis) between the 2-D peaks in common between a calibration standard data matrix, **N**, and "unknown" sample data matrix, **M**. In most cases the data matrices **M** and **N** that are being analyzed are small regions of separation data matrices which contain overlapping peaks. Regions of the 2-D separation data matrices in which the peaks are resolved are analyzed using standard peak height or volume methods.

The iterative alignment approach is based on the fact that the augmented data matrix [**M**|**N**], has a minimum rank when the 2-D peaks are aligned along the time axis in question. Hence, the alignment method shifts the 2-D peaks in **M** relative to **N** along the first column time axis until the resulting matrix [**M**|**N**] achieves a minimum rank. The alignment method uses the secondary eigenvalues from the singular value decomposition of [**M**|**N**] to find the peak shift correction associated with a minimum rank. Once the correct peak shift between **M** and **N** is determined, **M** can be aligned to **N**. This approach works fine for 2-D separations that have highly reproducible retention times along the second dimension time axis, such as is the case for most GC x GC data. Building upon our previous work, for 2-D separations that require shift corrections along both time axes, we report an alignment method that can be used to determine and correct the peak shifts along both time axes. By applying the original alignment method on the augmented matrix [**M**|**N**] and then subsequently on the augmented matrix $\begin{bmatrix} \mathbf{M} \\ \hline \mathbf{N} \end{bmatrix}$, peak shifts occurring along both time axes can be independently and successively corrected. In this work we have taken 2-D data collected from a GC x GC and have simulated 2-D data from other separation techniques by randomly adding retention time shifting that is consistent with reported precision from recent reports for LC x CE [10, 11]. We report how this improved algorithm corrects both dimensions in an independent, step-wise fashion.

4. Results and Discussion

In Figure 1 is shown the representative 2-D separation data for a four component mixture. Five replicate runs of the four component mixture served as sample runs. Next, a standard addition was made to the sample and run five times for the standard runs (not shown for brevity). In Figure 2 is shown the application of the improved alignment
algorithm to the five paired combinations of sample/standard runs. Note that in Figure 2 each data trace depicts one boundary contour line demarcating the peak boundaries for one run as in Figure 1. As shown in Figure 2, the improved alignment algorithm successfully shifted the five replicate 2-D sample runs in *both* dimensions. Subsequent GRAM analysis was very successful.

5. Acknowledgement

This work was funded in part by the Center for Process Analytical Chemistry (CPAC), a University/Industry Cooperative center at the University of Washington.

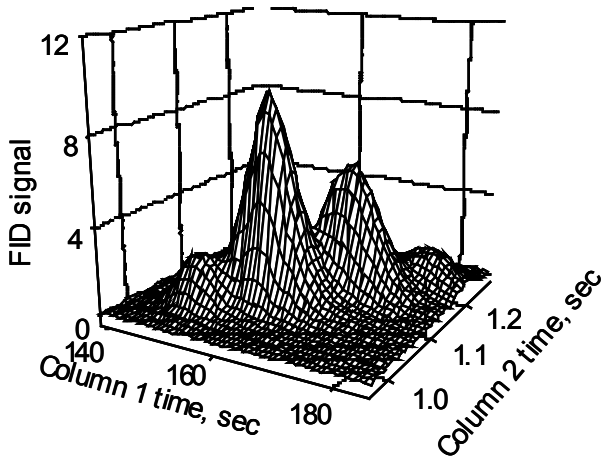

Figure 1. 2-D data obtained from run of four component mixture.

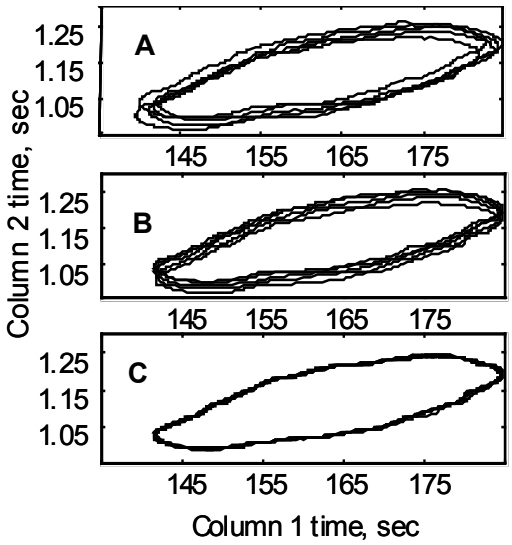

Figure 2. (A) Overlaid contour boundary plots (data as in Figure 1) from 5 runs, illustrating retention time variation on both dimensions. (B) After alignment along column 1 axis. (C) The 5 runs after alignment along column 2 axis.

6. References

1. C. A. Bruckner, B. J. Prazen and R. E. Synovec, *Anal. Chem.* Comprehensive Two-Dimensional Gas Chromatography with Chemometric Analysis **1998**, 70, 2796-2804.
2. C. G. Fraga, B. J. Prazen and R. E. Synovec, *Anal. Chem.* Comprehensive Two-Dimensional Gas Chromatography and Chemometrics for the High-Speed Quantitative Analysis of Aromatic Isomers in a Jet Fuel using the Standard Addition Method and an Objective Retention Time Alignment Algorithm **2000**, 72, 4154-4162.
3. C. G. Fraga, B. J. Prazen and R. E. Synovec, *J. High Resolut. Chromatogr.* Enhancing the Limit of Detection for Comprehensive Two-Dimensional Gas Chromatography (GC x GC) Data using Bilinear Chemometric Analysis **2000**, 3, 215-224.
4. C. G. Fraga, C. A. Bruckner and R. E. Synovec, *Anal. Chem.* Increasing the Number of Analyzable Peaks in Comprehensive Two-Dimensional Separations through Chemometrics **2001**, 73, 675 -683.
5. Z. Liu and M. L. Lee, *J. Microcol. Sep.* Comprehensive Two-Dimensional Separations Using Microcolumns **2000**, 12, 241-254.
6. J. B. Phillips, R. B. Gaines, J. Blomberg, F. W. M. van der Wielen, J. M. Dimandja, V. Green, J. Granger, D. Patterson, L. Racovalis, H. J. de Geus, J. de Boer, P. Haglund, J. Lipsky, V. Sinha and E. B. Ledford Jr., *J. High Resolut. Chromatogr.* A Robust Thermal Modulator for Comprehensive Two-Dimensional Gas Chromatography **1999**, 22, 3-10.
7. J. V. Seeley, F. Kramp and C. J. Hicks, *Anal. Chem.* Comprehensive Two-Dimensional Gas Chromatography via Differential Flow Modulation **2000**, 72, 4346-4352.
8. M. M. Bushey and J. W. Jorgenson, *Anal. Chem.* Automated Instrumentation for Comprehensive Two-Dimensional High-Performance Liquid Chromatography of Proteins **1990**, 62, 161-167.
9. M. M. Bushey and J. W. Jorgenson, *Anal. Chem.* Automated Instrumentation for Comprehensive Two-Dimensional High-Performance Liquid Chromatography/Capillary Zone Electrophoresis **1990**, 62, 978-984.
10. T. F. Hooker and J. W. Jorgenson, *Anal. Chem.* A Transparent Flow Gating Interface for the Coupling of Microcolumn LC with CZE in a Comprehensive Two-Dimensional System **1997**, 69, 4134 -4142.
11. A. V. Lemmo and J. W. Jorgenson, *Anal. Chem.* Transverse Flow Gating Interface for the

Coupling of Microcolumn-LC with CZE in a Comprehensive 2-Dimensional System **1993**, 65, 1576-1581.
12. E. Sánchez and B. R. Kowalski, *Anal. Chem.* Generalized Rank Annihilation Factor Analysis **1986**, 58, 496-499.
13. E. Sánchez, L. S. Ramos and B. R. Kowalski, *J. Chromatogr.* Generalized Rank Annihilation Method: I. Application to Liquid Chromatography-Diode Array Ultraviolet Detection Data **1987**, 152-164.
14. L. S. Ramos, E. Sánchez and B. R. Kowalski, *J. Chromatogr.* Generalized Rank Annihilation method II: Analysis of Bimodal Chromatographic Data. **1987**, 385, 165-180.
15. B. J. Prazen, R. E. Synovec and B. R. Kowalski, *Anal. Chem.* Standardization of Second-Order Chromatographic/Spectroscopic Data for Optimum Chemical Analysis **1998**, 70, 218-225.

Chemometric Movies Coming To a Theatre Near You

In our continuous effort to spread the word about the unbearable lightness of chemometrics, the Chemometric Society has plans to re-make classic moves in chemometric versions. These are among the movies we expect to appear on screen this year:

- ”SIMCAsablanca” starring Nouna & Svante
- ”For a few loadings more” starring Harald Martens
- “Honey I shrunk the regression coefficients” starring Sijmen de Jong
- ”Things to do at SSC7 when you’re drunk” starring ….*

* fill in whatever name you find appropriate