Fifty years of Chemometrics, fifty years with Chemometrics

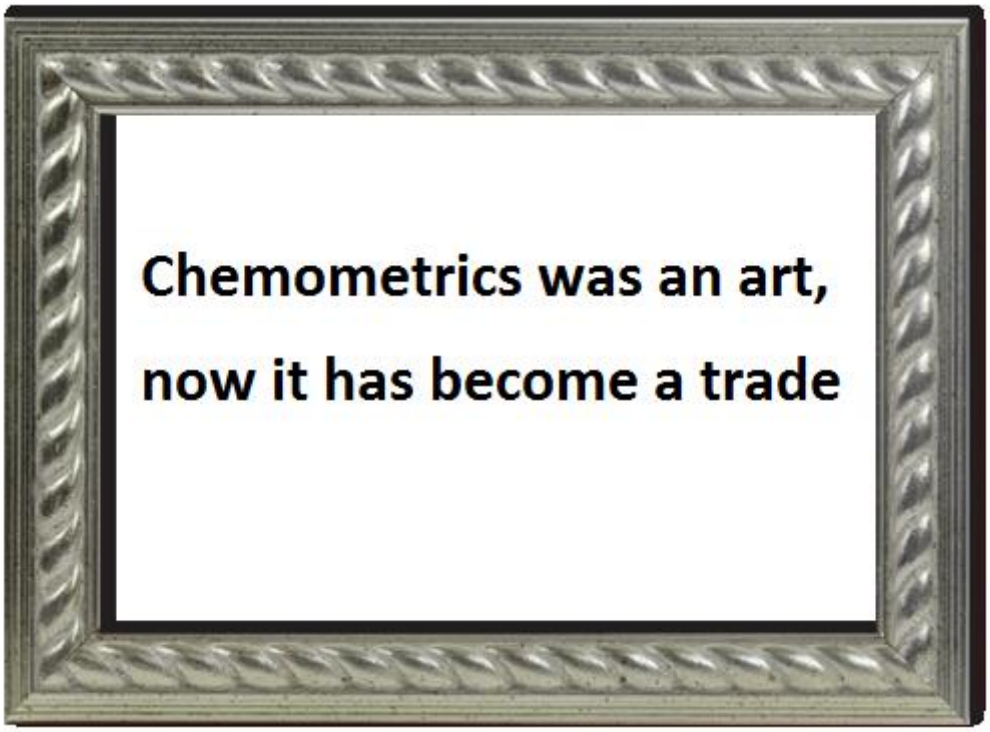
Research · September 2015

DOI: 10.13140/RG.2.1.2199.3445

1 author:

Michele Forina
Università degli Studi di Genova
157 PUBLICATIONS **2,119** CITATIONS

SEE PROFILE



In 2014, July, I was invited to write a paper for the Journal of Chemometrics (JC): “we are delighted to start inviting authors for a series of Perspectives under the temptative heading "The pioneers in Chemometrics series" and we would be honoured if you would accept to be among the selected few contributors. Your commentary on the origins and the future of the field”.

In December I sent a first version to the Editor.

In 2015, March, I sent a new version, modified according to the suggestions of the Editor. He answered: “This looks great and ready to be submitted. Please proceed using the web system at: https://mc.manuscriptcentral.com/cem. Please pay attention to mention in the letter to Paul Gemperline (Editor) that your contribution is an invited feature article in the series 'The pionneers in chemometrics series".

In May I received the evaluation of Referees:

Reviewer: 1

“I performed an expedient English fix although some sentences may still need a further polishing which can be done at the next stage. There are a number of questions in the comments of the attached doc. These attempt to make the paper more understandable.”

Reviewer: 2

“Michele Forina is a pioneer in the field of chemometrics and a well respected chemometrician, therefore a historical perspective from this author is potentially interesting to readers of this journal. The present submission, unfortunately, is primarily a historical account of scientific meetings and persons the author met at these respective meetings without much insight into how the field developed. With all due respect to Prof. Forina, I do not recommend publication of this perspective in its present form.”

The Editor suggested me to revise the manuscript:” I sincerely think that starting again from reviewers 1 version of the paper and pointing on some scientific facts and algorithm you participated to and develop, the reviewers requirements could be met. You should know as well

that you are not the only founders author experiencing these kind of difficulties. It turned out to be a quite difficult challenge to produce these papers, mixing personal, historical and scientific facts."
So, in June, I submitted to the Editor a new version of the manuscript, with attention to the remarks of Reviewer 2, because I'm used to respect the reviewers.
Because of some remarks of the Editor, I submitted in July the version number 4.
It returned me as version number 5, ready to be submitted to the JC, to have new remarks, useful to prepare the nest versions.
Many things bother me.
I have to change from version 1 to version 2 "the positive and the negative side of Chemometrics" in "the pros and the cons of Chemometrics". Not exactly the same significance. Then the Editor suggested me to put the pros and the cons in a Table. Finally, in version 5, the "pros and cons" disappeared and they were substituted by "the positive and the negative". The loops are a very important characteristic of algorithms, but in this case they offer also other possibilities, useful for many new versions, as the top and the bottom, the white and the grey, the smell and the stench, and so on.
Moreover, many parts were cancelled, or shifted, or modified.
So I decided to stop to write new versions, and the manuscript below, from version 3 with some elements of version 1, is presented to my followers in Research Gate.

2015, September

Fifty years of Chemometrics, fifty years with Chemometrics

Michele Forina

A personal history of a life in Chemometrics

Each important step of this personal history corresponds with an evolution in my research frequently with the collaboration with other chemometricians. So this personal history is not so much my own story but is part of the history of many chemometricians practicing today.

About fifty years ago, a few enthusiastic researchers began to work on what today we call Chemometrics. They were, with few exceptions, analytical chemists, as I was, with some background in statistics or mathematics.

Two papers by an unknown young scientist from Washington State University regarding linear learning machines [1] and the best delimiter for least squares pattern classifier [2] were my first contact with multivariate data analysis. The paper about linear learning machines (the Perceptron) is, in my knowledge, the first example of use of neural nets in chemistry.

In 1974 I received the letter demonstrate here as Figure 1 announcing the birth of the International Chemical Society (ICS). I became a Chemometrician (at the lowest level). After some weeks, I received ARTHUR, the executable and the source. The latter was an inexhaustible reservoir of algorithms, in spite of the difficult FORTRAN source decoding, wherein all of the matrices are merged in one vector and appropriately addressed.

To compute the principal components (the Karhunen-Loève projection of ARTHUR), it was necessary to use four ARTHUR modules, SCALE for autoscaling, KAPRIN to compute loadings, KATRAN to compute scores, and VARVAR, to obtain a superb plot, as shown in Figure 2.

To improve both my knowledge of the algorithms and the efficiency of these analytical tools, I began to compose some computer programs for my new computers; an Olivetti 101 (the first world personal computer), followed by the Olivetti P6060.

I baptized my computer modules with the name PARVUS, as they were a light imitation of the giant ARTHUR. The development of PARVUS has been my main focus henceforth, and presently I'm refining the latest version to be distributed free very soon (I hope).

In 1978 I was visited by Sergio Clementi, an organic chemist from the University of Perugia who had first come in contact with Chemometrics in a meeting in Amsterdam. Sergio's interest was directed to LFER (Linear Free Energy Relationships) and so to the applications of Chemometrics in organic chemistry. Really the International Chemometrics Society had at the beginning two sections, one for analytical chemistry, the second for organic chemistry.

Dear Prospective Chemometrician:

Although statistics has been a tool of the chemist for many years, the recent literature shows a substantial increase in the number of novel applications of statistics and non-statistical mathematics to problems in the field of chemistry. The reasons for this increase are many, and include such things as increasing amounts of quantitative data produced in all branches of chemistry, greater access to computers and difficulties for theory to describe data as more complex problems are attacked. While many of the statistical and mathematical methods are known to all, new and powerful methodology has permeated chemical applications from such fields or subfields as: estimation theory, decision theory, pattern recognition, information theory, optimization, artificial intelligence, spectral and wave form analysis, numerical analysis, cybernetics, and many others.....

On June 10, 1974, the Chemometrics Society was begun. So far, it is an informal society and its primary function is communication. The purpose of this letter is to invite you to join the Chemometrics Society and participate in the communication of mathematical and statistical concepts and applications in the field of chemistry. There is little doubt that there are many societies that a chemist may choose to join, and it is a sad fact that these societies usually end up with the member serving the society and with very little return. It is the purpose of this Society to serve its membership and to ask for very little in return. The Chemometrics Society will exist mainly as a special interest group for the communication of research ideas among its members. We ask only that you keep the society informed of your research publications and in return we will make a directory of members and research publications available to the members. It is hoped that authors of papers will not only communicate paper titles and journals published, but also a short summary, if not a preprint or reprint of the paper itself, so that a scan of literature will be greatly facilitated. There will be no dues for this service as, if each member carries his own load, the costs will be kept to a minimum.

Our first idea is to publish a newsletter containing the membership list and a summary of what we have learned from responses to this letter. Therefore, we ask for your comments, suggestions, and most importantly, your indication of interest in the society. Please send your full name, address, telephone number. It would be very helpful if you could include a bibliography of past and current papers and manuscripts that are in for publication. In addition, we would appreciate your notifying us about prospective members among your colleagues. If all the members cooperate in this endeavor, the first newsletter will be a valuable aid for anyone involved in the application of statistical and mathematical methods in chemistry.

Hoping to hear from you soon

For the Chemometric Society, yours faithfully

Bruce R. Kowalski
Laboratory for Chemometrics
Department of Chemistry
University of Washington
Seattle, Washington 98195

Svante Wold
Research Group for Chemometrics
Institute of Chemistry
Umea University
Sweden

Figure 1 – The letter announcing the birth of the Chemometrics Society

Figure 2 – A PC plot of ARTHUR

In 1980 Bruce Kowalski visited me in Genova. At that time he was on sabbatical at Garching University. Bruce's presence was illuminating, not only for the transmission of knowledge but also for his qualities as a generous man, teacher and organizer. The conferences involving Bruce were a continuous stimuli. He was accustomed to speaking in a manner that transmitted his ideas clearly by all intellectual levels and never exercised superiority (as many lecturers do).

In April 1981 together with Luc Massart we organized the first Italian meeting of Chemometrics. Luc was the second most influential person in my history. I found to be akin to Bruce but with a very different style, he was an exceptional lecturer with clarity, organization, and the ability to demand attention. Luc came to Italy many times and his lectures have sculpted Italian Chemometrics.

In 1982, thanks to Luc, I was invited to give a lecture at the Food Research and Data Analysis meeting held in Oslo. The meeting gathered experts in the areas of food, statistics, psychometrics and chemometrics together for the first time.

There were many important people at the Voksenasen Hotel seat of the meeting. Some of them played a very important role in my life and from them I learned many things. I remember Herman Wold, with his wife

Bianca (the last daughter of Svante Arrhenius), Karl Norris, Harald Martens, Karoly Kaffka, the very young Paul Geladi and Tormod Naes.

And Svante Wold.

Svante has been, together with Bruce and Luc, the third pillar of my life in Chemometrics.

What I admire in Svante (among his many qualities) is the clarity of his research papers. The algorithms of important techniques, as NIPALS, SIMCA, PLS, are described very carefully making them easy for me reproduce in detail. This was an important element for PARVUS. The detailed knowledge and the full understanding of the algorithms are very important for a chemometrician. I recall a referee comment on a paper I had submitted on wavelets affirmed that the Mallat pyramid is applicable only to the Haar filter. The referee, evidently an "expert" of wavelets, had no knowledge of the lift algorithm, the base of the wavelet decomposition for all the wavelet filters.

In the Oslo meeting Harald Martens explained the principal components by means of a giant inflatable banana. Harald gave me the banana, which is now in my department, entrusted to my younger coworker, Paolo Oliveri.

At the end of the meeting, Svante and Harald presented other sketches.

These examples are collected in the supplementary material of this paper, in a revised form with other examples used to teach Chemometrics and explain multivariate statistics in more exciting and accessible ways:

S. Wold - The four friends

S. Wold and H. Martens - The old manuscript

M. Forina – In the beginning

X. Tomas - The elephant.

The lectures of the Oslo meeting are collected in a book [3]. I presented a lecture on the classification of olive oils, with the use of many classification techniques, from KNN to Quadratic Discriminant Analysis to SIMCA. The results with SIMCA were not so good. At lunch I heard by chance a comment of Svante to Forrest Young (the famous psychometrician): "I'm not responsible for these results, because he didn't use my software". The next morning at breakfast I received the first lesson of Svante.

In 1983 Harald Martens at the Italian Conference of Analytical Chemistry in Parma began his lecture playing a piece of music with only one guitar string. He said: "This is univariate statistics". Then he played the same piece using all the guitar strings. He said: "… and this is multivariate statistics".

Harald was just traveling to Cosenza, for the NATO Advanced Study Institute (ASI) meeting on Chemometrics. The Advanced Study Institute was promoted by Bruce and Luc. I was charged of the local organization with Sergio Clementi and Giovanni Latorre.

The NATO ASI was an exciting two weeks experience. The lectures from which have been collected in a book [4], wherein the seeds of the next twenty years of Chemometrics were sown. Almost all of the pioneering

chemometricians attended ASI were in Cosenza, Bruce, Luc, Svante, Harald, Bernard Vandeginste, Kurt Varmuza, Max Feinberg, Stanley Deming, Ivar Ugi, Gabor Veress, Kim Esbensen, Pierre Van Espen, Paul Levi, Leonard Kaufman, Steve Brown, Lloyd Currie, Edmund Malinowski, William Dunn, Stuart Hunter and William Hunter. There were also young up and coming characters of the field: David Veltkamp, Tormod Naes, Ann Smeyers, and Lutgarde Buydens.

Cosenza was the first time that I presented PARVUS internationally, with some results obtained from the characterization of typical foods such as olive oil and wines. A few months after I wrote a short presentation of PARVUS on TRAC (Figure 3). It was published as "PARVUS: an extendable package of programs for data exploration, classification and correlation", ESS Elsevier Scientific Software, Amsterdam, 1988

Computer Corner

PARVUS

In this contribution, Prof. Forina describes a package for pattern recognition, which to my knowledge must be the most complete such package written directly for microcomputers. It contains almost all display methods and supervised learning methods currently used by analytical chemometrists. Professor Forina was recently appointed President of the Chemometrics Society. He succeeds the President-Founder of the Society, Prof. Kowalski from Seattle.

D. L. MASSART

Figure 3 – PARVUS presentation in TRAC

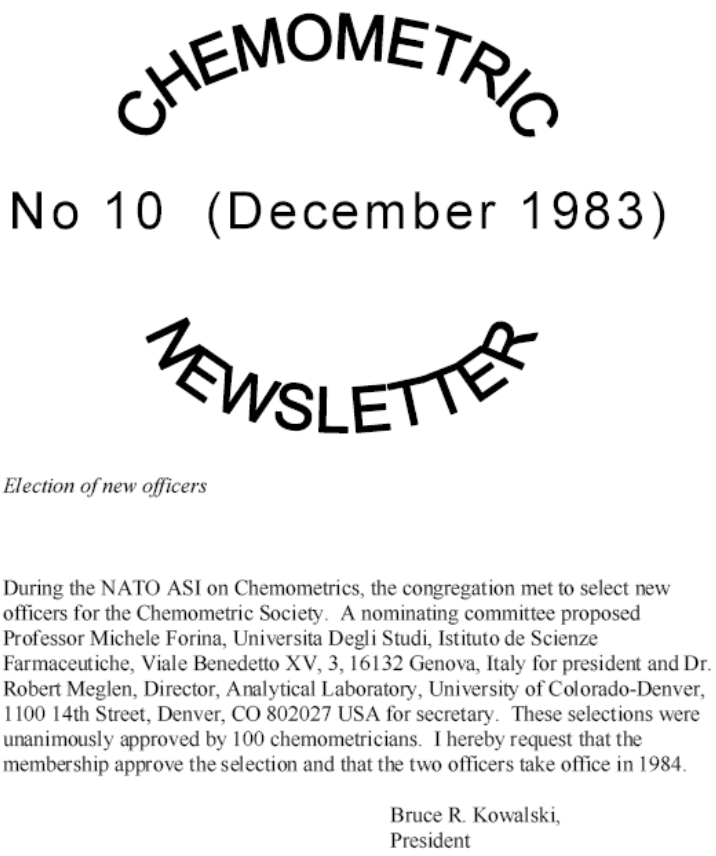
During my lecture in Cosenza I presented results obtained with SIMCA, good results. I used a modified version of SIMCA. The original SIMCA algorithm computes the range of the scores on the significant principal components of the studied class. Then the class model is obtained by increasing the range. I used instead a reduced range. Moreover I redefined the SIMCA box using a modified statistics. The next day Svante Wold showed the modified SIMCA box in his lecture, and he added the flexibility to the other qualities of SIMCA.

Generally people use SIMCA almost as a black box. The quality that SIMCA can work also when the number of variables is more than that of the objects is frequently interpreted as the possibility to use hundreds of variables, possibly noise variables. As the noise increases it reaches the first components, and the model become useless.

At the end of the Institute meeting, I was appointed president of the Chemometrics Society (Figure 4) following ten year's of Bruce's presidency.

In 1986, May, I organized the third Chemometrics in Analytical Chemistry (CAC) conference in Lerici, a small village on the east ligurian coast, made famous by Byron and Shelley. Previous CAC meetings had taken place

in Amsterdam in 1978, and Petten. Following Lerici, CAC became the most important international conference of chemometricians.



CHEMOMETRIC

No 10 (December 1983)

NEWSLETTER

Election of new officers

During the NATO ASI on Chemometrics, the congregation met to select new officers for the Chemometric Society. A nominating committee proposed Professor Michele Forina, Universita Degli Studi, Istituto de Scienze Farmaceutiche, Viale Benedetto XV, 3, 16132 Genova, Italy for president and Dr. Robert Meglen, Director, Analytical Laboratory, University of Colorado-Denver, 1100 14th Street, Denver, CO 802027 USA for secretary. These selections were unanimously approved by 100 chemometricians. I hereby request that the membership approve the selection and that the two officers take office in 1984.

Bruce R. Kowalski,
President

Figure 4 – Chemometrics newsletter of December 1983

During the conference, seven Spanish participants founded the Spanish section of the Chemometrics society, with Enric Casassas as president. Together with two of the Spanish Cheometric Society members, Jesus Lopez Palacios and Luis Sarabia of Burgos, I initiated an Erasmus net for Chemometrics. Later the net enlarged to encompass thirteen universities in Italy, Spain and France. Many students migrated into this network to improve their knowledge of Chemometrics, many of them are now teachers or researchers in our field. The net was also a very efficient vehicle to stimulate collaboration and nurture of friendships.

In Lerici, Luc Massart invited me to participate, in collaboration with Christian Ducauze and Max Feinberg of Paris, and Roger Phan-tan-luu of Marseille, into a European Comett pilot project for teaching Chemometrics called Eurochemometrics. This project was realized in the years from 1987 to 1989, with three international schools, the first in Eguilles, near Aix-en-Provence, the second in the Tortosa Castle, near the mouth of Ebro river, the third in Gargnano, on the shores of Lake Maggiore. These schools have also been an important medium for the dissemination of Chemometrics. In a second phase, Eurochemometrics extended to other countries with an increasing number of schools. My team continued to organize schools in Italy between 1990 and 2000, supported by Luc, Johanna Smeyers, Lutgarde Buydens, Roger Phan-tan-luu, Michelle Sergent, Luis Sarabia, Rafael Cela, and Jure Zupan.

In preparation of the first Eurochemometrics School, I had the first meeting with Roger Phan-tan–luu and the delicious Michelle Sergent. Roger has been important to me, in much the same way as Bruce, Luc and Svante, in research, teaching and friendship. His incomparable didactic ability combined with expert experimental design were crucial for the dispersal of the scientific methodology in Italy and Spain. Before meeting Roger, I had a pale idea of experimental design (from the lectures of the two Hunters and of Deming in Cosenza). After Roger, the level of my research, particularly in the characterization of typical foods, increased significantly.

In Lerici, the participants designated candidates for the presidency of the international Chemometrics Society. Luc Massart was appointed president for three years, starting with 1987. Wolfhard Wegscheider was the new secretary. Wegscheider moved from Graz to Leoben, Luc had too many engagements with research, teaching, writing books and organizing projects so the nomination of the fourth president was delayed. Even today, the website of the North American Chapter of the International Chemometrics Society indicates Luc as president and Wolfhard as secretary of the Society.

During the last evening in Lerici, I was in a restaurant with some friends. Among them, Luis Sarabia and Max Feinberg. As we were used to talking in our native language the idea was born to organize a Latin conference on Chemometrics enabling Latin people the ability to discuss Chemometrics in their native languages. This was the origin of the Colloquium Chemiometricum Mediterraneum. The first Colloquium was held in Barcelona, November 1987, organized by Enric Casassas. Subsequent meetings were held in San Miniato (Italy), Bastia (France), Burgos (Spain), Ustica Island (Italy), St. Maximin (France), Granada (Spain) and Bevagna (Italy). In the history of the Colloquium I remember only one communication in true Latin, in St. Maximin: the speaker was Rolf Carlson: “De experimentis combinatoriis in synthesibus organicis et de proiectionibus ad vectores latentes“. In more recent years the Colloquium has become a mainly English transmitted event with exposure to the English language increasing among Latin scholars.

So, after Lerici and for many years to follow I picked the fruits. I continued in research, in teaching and in the other usual activities of a university professor. A pleasant life, because, as Roger Phan-tan-luu frequently remarked: “We are not working, we’re having fun”. Work is the ensemble of activities in administration, exams …

Chemometrics was growing rapidly. The new journals, Journal of Chemometrics and Chemometrics and intelligent laboratory systems, continuously published papers on new tools and new application fields. The atmosphere was extremely challenging, an invitation to emulation.

I had two main areas of interest, the characterization of typical foods, mainly wines, olive oil and cheese, and the improvement of my software, PARVUS.

I remember here some points of my work, those I think more representative.

To characterize foods I applied classification methods, then class modeling techniques. So I studied these techniques. UNEQ [5], one of the most important class-modelling techniques, appeared in 1986. To estimate

the boundary of the class model UNEQ used the Defrise-Gussenhoven statistics, a correction of the χ^2 statistics. Because of some anomalies in the results (negative distance for objects close to the class centroid), I decided, with Luis Sarabia, to study the distribution of the distances by means of Montecarlo experiments. The results [6] showed that the correct statistics to be applied are the β statistics for the objects in the evaluation set and the T^2 statistics for the objects in the evaluation set. The use of χ^2 statistics enlarges the model, so that some outliers in the training set can't be detected. Also the use of the Defrise-Gussenhoven correction enlarges the model and reduced the specificity.

Really at that time many chemometricians had a limited knowledge of statistics, and also statisticians rarely had experience with multivariate statistics. So, both the Authors of UNEQ and we had no knowledge that the principles of UNEQ with the T^2 statistics for the objects in the evaluation set were presented by Harold Hotelling in 1947.

In the same time, I studied the potential functions classification technique, applied by Danny Coomans in Chemometrics in 1982. The objective of this study [7] was to use potential functions as a class modeling technique, in that similar to SIMCA and UNEQ but suitable for complex distributions. One of the procedures used to compute the boundaries of the class model, that called "the equivalent determinant", is what I think my best idea in the application of statistics. Figure 5 shows an example of application, with a data set used frequently to evaluate the performances of artificial neural nets. What a pity that potential functions methods are rarely used, probably because they are not present in the most important commercial software, because they have many interesting characteristics, as the possibility to build a class space and to compute from the probability density a distance equivalent to the distance from the center of the class model of UNEQ and from the SIMCA model.

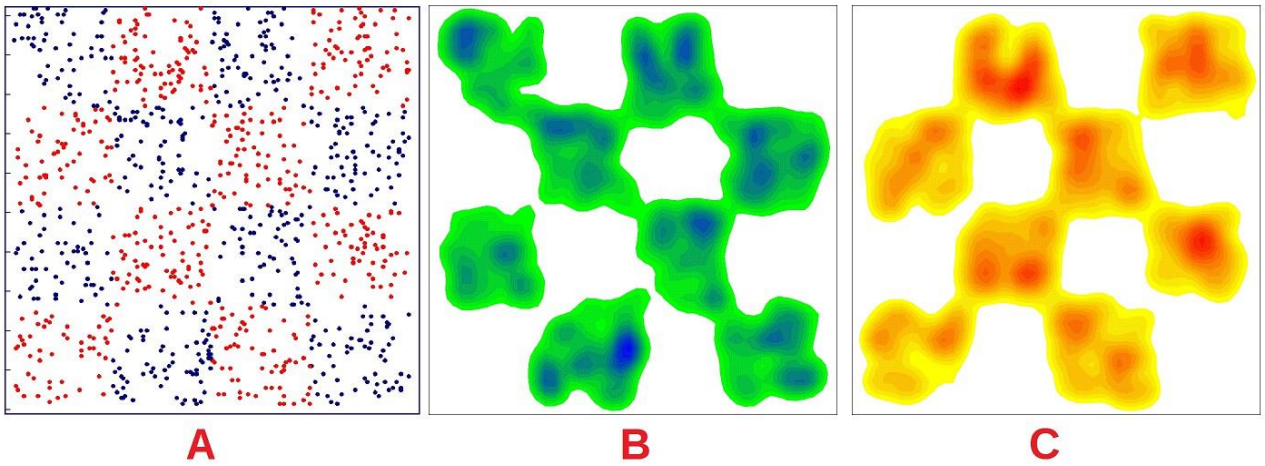


Figure 5 – The data set Checker, with two categories (A). The 95% class models and the probability densities computed by means of the potential functions for the first (B) and the second (C) category.

We (me and my coworkers) used potential functions also in sampling design [8]. In the characterization of Liguria olive oil, we had the problem of too many samples. To obtain a balanced data set, first we selected

some objects by means of the Kennard-Stone design, to cover uniformly the main climate characteristics. Then we used an algorithm based on potential functions to select other samples to represent the density of production.

In that regards the characterization of typical foods, we began [3-4] with a data set, Oliveoil, where our knowledge of important details (exact location of olive groves, year of production, storage time and conditions, pressing technique, …) was very poor. With the time we realized that the samples must represent all the variability factors. Among all the studies appeared on typical foods, almost never the effect of storage was considered. We studied this effect during all the time from the production until the expiry date. Figure 6 shows the result of this study on the visible spectra transformed by means of row profiles where the differences in saturation are deleted and those in tonality retained. The effect of the storage time is significant, but less than the effect of the other factors, that explains the result in Figure 6A.

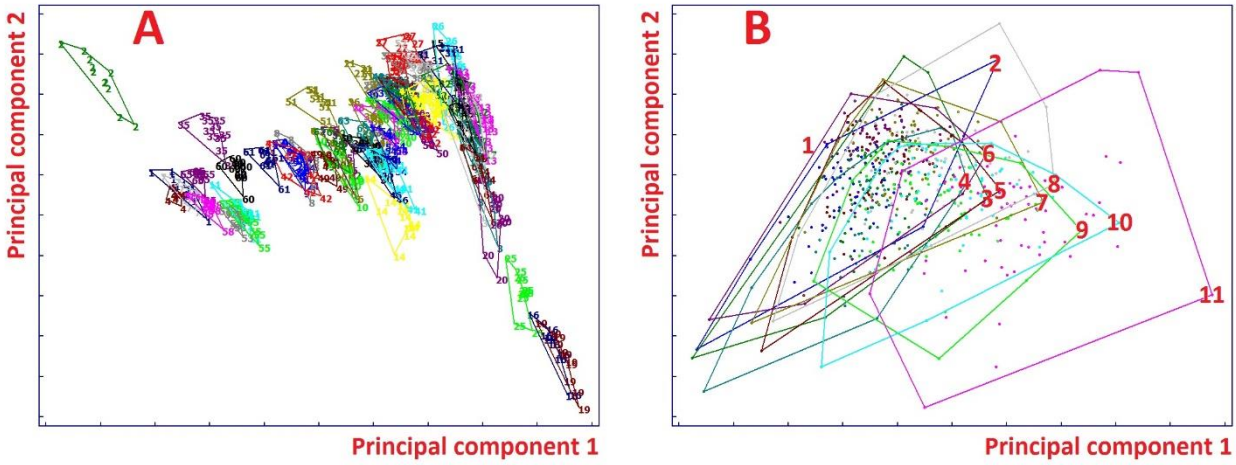


Figure 6 – PCA of 65 samples of olive oil. The visible spectra were recorded 11 times, from the arrival in the laboratory to the expiry date. A: PC analysis of row profiles. The 65 convex hulls show the dispersion of each oil sample during the storage time. B: differences between the spectra and the first spectrum (1). The convex hulls enclose the 65 oil samples.

We studied also the clustering techniques [9]. Here the user interprets the dendrogram paying attention to both axes with the instinctive sensation that close objects are similar, that can be true but also false. The dendrogram is based on the matrix of similarities, and Robinson [10] suggested to change the original order of the objects to obtain a similarity matrix as close as possible to a “Robinson matrix” where the magnitude of the elements monotonically decreases moving away from the diagonal. The Robinson procedure is called seriation and it is specially used in archaeometry, to put the objects in chronological order.

The previous strategies to obtain a Robinson matrix work directly on the similarity matrix, with the similarity defined by equation:

$$s_{ij} = 1 - \frac{d_{ij}}{d_{MAX}}$$

where d_{ij} is the distance between the two objects and d_{MAX} is the maximum distance between two objects in the data set.

With N objects there are 2^{N-2} possible different orders of the objects. With the example used here, the 50 objects of the class Setosa of the data set Iris used by Ronald Fisher presenting linear discriminant analysis, the number of possible orders is about $3\ 10^{14}$, too many.

Our seriation algorithm (DSA: Direct Seriation Algorithm) directly works on the dendrogram. DSA defines the "similarity function" as:

$$SF = \sum_{i=1}^{N-1} \text{similarity}\ (i, i+1)$$

where the order of the objects (index i) is the order of the abscissa of the dendrogram. A small value of the similarity function means that there are some adjacent objects with small similarity.

The agglomeration procedure of clustering partially orders the objects. So, the order after clustering seems a good starting point for a seriation algorithm. In the case of the data set Setosa, SF is 31.5 for the original data. The worst order corresponds to SF = 23 about. The best order (obtained with many billions of permutations, two days computer time) is 40. After agglomeration (single linkage) SF is 35. DSA results in SF 39.

The algorithm shown below performs all the possible combinations of translations of branches and (if required) rotations in a selected small interval of branches, parameter DEEP (the branches are ordered according the similarity of agglomeration). The number of these combinations is 2^{DEEP}-1 for translations. The number of combinations for rotations is less, because the rotation around branch N-1 simply inverts the order of all the objects, and the rotation around a branch that connects only two objects corresponds exactly to the translation.

The DSA algorithm is presented below, for the case of only translations. When also rotations are required, the loop B is repeated with rotations instead of translations and with FIRSTBRANCH = N – 2.

In the case of a large number of objects, translation and rotation of branches can be limited to the high level branches, because the low level branches join very similar objects. The parameter LIMIT indicates the lowest branch for corrections.

```
DSA begin
    Select DEEP interval, from DEEPMINIMUM (suggested 2) to DEEPMAXIMUM (suggested 10)
    Select LIMIT
    OPTIMUM = starting value of the objective function SF
    N = number of objects
    FIRSTBRANCH = N-1
```

```
CORRECTIONS = 0
    A) FOR DEEP = DEEPMINIMUM to DEEPMAXIMUM (
            B) FOR CYCLE = 1 to DEEP -1
               CYCLECORRECTIONS = 0
                    C) FOR BRANCH = FIRSTBRANCH + CYCLE -1 To DEEP + LIMIT step -DEEP.
                            LOCALOPTIMUM = OPTIMUM
                                    D) For all the 2^DEEP-1 possible combinations of branches in the interval
                                    BRANCH – BRANCH + DEEP - 1 transpose around the selected
                                    branches and compute SF.
                                    If SF > LOCALOPTIMUM then
                                            LOCALOPTIMUM = SF
                                            Store the combination as OPTIMUMCOMBINATION
                                    End if
                                    Cancel the transpositions corresponding to the combination.
                                    END FOR D
                            If LOCALOPTIMUM > OPTIMUM then
                                    CORRECTIONS = CORRECTIONS + 1
                                    CYCLECORRECTIONS = CYCLECORRECTIONS + 1
                                    Modify dendrogram according to OPTIMUMCOMBINATION
                            END IF
                    END FOR C (next BRANCH)
                    IF CYCLECORRECTIONS = 0 then next DEEP
            END FOR B (next CYCLE)
    END FOR A (next DEEP)
DSA END
```

Figure 7 shows the color-coded representation of the matrix of similarities between the 50 objects of the class Setosa, before and after seriation. In twelve seconds the algorithm evaluated about 15000 rotations or translations.

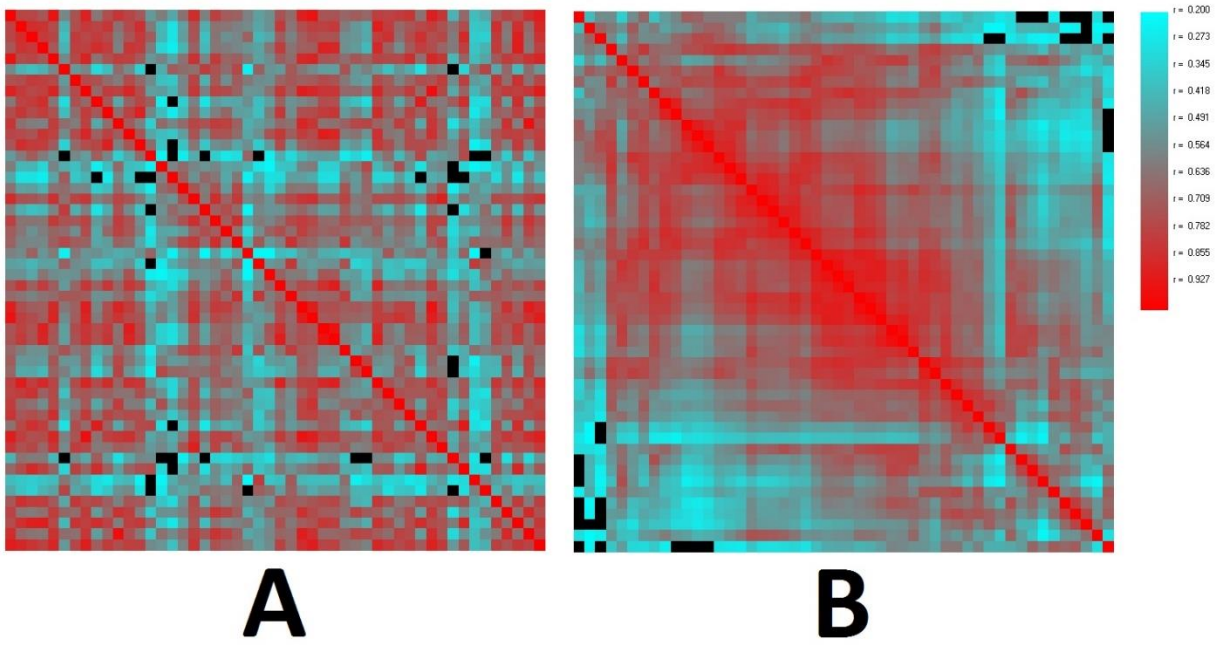


Figure 7 – Similarity matrix of the 50 objects of class Setosa. A: initial order, B: after seriation

I worked also with regression problems. My first contact with PLS was in the Oslo conference [3], just the presentation of PLS to the scientific community. Working with PLS and NIR spectra, I had the problem of noise, because a large part of the spectra were useless. So we try to develop methods for the elimination of useless predictors, as many colleagues were doing (as Massart with UVE, Clementi with GOLPE, Lindgren with IVS, Martens with a cutoff value for the PLS weights).

IPW (iterative predictor weighting) is based on the fact that the PLS weights that define the direction of the latent variables are computed by the covariance predictor-response, and uses the importance of the predictors

$$z_v = \frac{|b_v| \; s_v}{\sum_{v=1}^{V} |b_v| \; s_v}$$

where b are the coefficients of the closed form and s are the standard deviations of the predictors.

The IPW algorithm contains the classical PLS algorithm. It is:

[a] Set: $\mathbf{Z} = \mathbf{I}$ (Identity matrix)

[b] Set: **X**: Original matrix of predictors, **y**: Original vector of response

[c] Scale predictors and response

[d] Multiply matrix of predictors by the diagonal matrix of importances:

$$\mathbf{X} \Rightarrow \mathbf{X} \; \mathbf{Z}$$

[e] $\mathbf{w}^T = \mathbf{y}^T \mathbf{X} / \mathbf{y}^T \mathbf{y}$ PLS weights

[f] $\mathbf{w} = \mathbf{w} / \|\mathbf{w}\|$ Normalization of PLS weights

[g] $\mathbf{t} = \mathbf{X} \mathbf{w}$

[h] $c = \mathbf{t}^T \mathbf{y} / \mathbf{t}^T \mathbf{t}$

[i] $\mathbf{p}^T = \mathbf{t}^T \mathbf{X} / \mathbf{t}^T \mathbf{t}$

[j] $\mathbf{X} \Rightarrow \mathbf{X} - \mathbf{t} \, \mathbf{p}^T$

[k] $\mathbf{y} \Rightarrow \mathbf{y} - c \; \mathbf{t}$

Go to step [e] to compute the next PLS latent variable. The complexity of the model is obtained by predictive optimisation.

[l] Compute **Z** with the significant number of PLS components (the regression coefficients **b** are referred to the original predictors)

[m] If required, delete predictors with importance less a cut-off value. Recompute **Z**.
Go to step [d] for the next IPW cycle.

Because of the multiplication in step d, the predictors with small importance will have in the next cycle a lower weight and generally their importance will further decrease until disappear.

We worked many times with IPW in real problems, but always simultaneously taking into account other methods.

The end of my activity was approaching when we developed a new class-modeling technique, CAMM [12]. The motive was my dislike for the artificial nets classifier. The examples used to demonstrate the power of some nets indicate clearly that simple transforms of the original variables can change the system from non-

linearly separable to linearly separable. Simply, we added to the original variables some new variables, the distances from the class centroids, the leverages or the Mahalanobis distances, also the SIMCA distances. This approach is very efficient in many apparently complex cases. CAMM can't be applied to data sets similar to that used above (Figure 5) for the potential functions classifiers, but very probably real data are not so complex.

The years passed quickly and in October 2010 I retired.
Now I spend almost half of the year in Pietrasanta, in the plain between the Apuan Alps and the sea. I continue to work in Chemometrics completing a book [13], working on the continuous revision of PARVUS, while studying some of the old and new techniques of Chemometrics.
However, slowly, I'm returning to the solitude from which I began to escape about four decades ago.
This because many great masters or good friends, top figures of Chemometrics, passed away.

William Hunter December 1986
Herman Wold February 1992
Jean Clerc June 1998
Enric Casassas February 2000
Ivar Ugi September 2005
Luc Massart December 2005
Leonardo Lampugnani – August 2007 (Leonardo organized the first Colloquium in Italy)
Mario Castino September 2009 (Mario was the first to apply LDA to the classification of Italian wines, in 1973)
Sijmen de Jong October 2010
Paul Lewi August 2012
Bruce Kowalsky December 2012
Gerrit Kateman March 2013
Karoly Kaffka June 2014

In my solitude, the memory of times spent together, of what they did for me, sometimes of what we did together, continues to be a good company.

Some additional note about the history of Chemometrics

Svante Wold was the first to use the word Chemometrics, written in Swedish in 1972 (Forskningsgruppen för Kemometri). However before 1970 many papers were published with a neat chemometrics content, as those cited here [1-2].

More, our pride in being chemometricians has far origins.

Many years ago, William Gosset entered New College Oxford. He obtained a First Class degree in analytical chemistry in 1899. Gosset obtained a post as a chemist with Arthur Guinness Son and Company in 1899, working in quality control (the humble work of many analytical chemists). Working in the Guinness brewery in Dublin he had to use statistics. Because the results obtained with the normal distribution, the "law of error" were not satisfactory, he studied more and in 1904 presented the results of his studies as "The application of the "law of error", in an internal report. In 1904 he published a paper in Biometrika, VI (1908) p. 1, entitled "The probable error of a mean". In this paper he introduced a new statistics that we know with the name used by William Gosset to sign the paper, Student.

A second great scientist that I like to recall is Harold Hotelling.

His T^2 statistics are the base of many Chemometrics techniques, not only UNEQ.

Hotelling introduced his T^2 statistics for application in the multivariate quality control and he was also the father of the canonical correlation analysis, the first technique at the fourth level of patter recognition and a pioneer of experimental design. I came to know the Hotelling twin-pan balance by means of Roger Phan-tan-luu. I used Roger's development to explain experimental design many times. Finally, I would like to highlight a quote from Hotelling c.1940:

"Qualifications of a good teacher of statistics include, first and foremost, a thorough knowledge of the subject. This statement seems trivial, but it has been ignored in such a way as to bring about the present unfortunate situation".

When we change "statistics" in "Chemometrics" we obtain, unfortunately, a very actual statement.

Hotelling's presentation (the balance) can be found in the supplementary material.

I was invited in Umea c.1995 to be the opponent for Fredrik Lindgren's thesis. It was requested that I illustrate the thesis in an accessible way for non-specialist observers who were present at the defense. Shockingly, the conference room was full with more than one hundred people! I used the second example in the supplementary material to introduce PLS to this heterogeneous people. In hindsight, however it was probably only the chemometricians who understood!

The use and misuse of Chemometrics

On the positive side, Chemometrics is the chemical discipline that uses mathematical, statistical and other methods:

1. to design or select optimal measurement procedures and experiments, and
2. to provide maximum relevant chemical information by measured or computed data.

Note that the definition excludes the term "multivariate". Chemometrics also uses univariate statistics, when it is sufficient. The important thing is acquiring the relevant information in whichever form.

"Data" can mean chemical quantitative, as concentrations, qualitative ordinal as trace-minor-major constituent, physical, as absorbance, or computed as the interaction energies used in QSAR.

Chemometricians can obviously invent new methods, or modify old methods. They can sometimes use artificial data. A chemometrician must always remember that Chemometrics is a chemical discipline and the objective is paramount. In the above definition, the first step (design or select …) is in function of the second: to obtain relevant information about a chemical problem.

On the negative side, Chemometrics is the discipline that fabricates new methods or cocktails of methods to extract useless information from artificial data.

I remember a paper were a transform based on the curvature function of Riemannian warped space was presented. It was a very original strategy with appealing words but contained too many weak points.

Riemann (integration) appears in 2001 in a different paper (different author) where a number of artificial objects are added to the original ones, means of two objects of the same class. The addition of artificial objects continues adding the means of the original objects and of the artificial objects of the first generation. Then the original objects are classified by KNN taking into account the artificial objects. The method is a powerful classification tool and has been applied to a very important problem: the classification of women in a town by means of the metal content in their breast milk. In the same paper some figures show the same quantity on both axes, with an increased range, probably to have space for remarks. The mysterious quantities (ANN t_1 and t_2) have an exceptional accuracy, so that they are expressed with eight significant digits (Figure 8). It is a shame that both methods have only used by their author and that nobody has deepened the research on the breast milk problem.

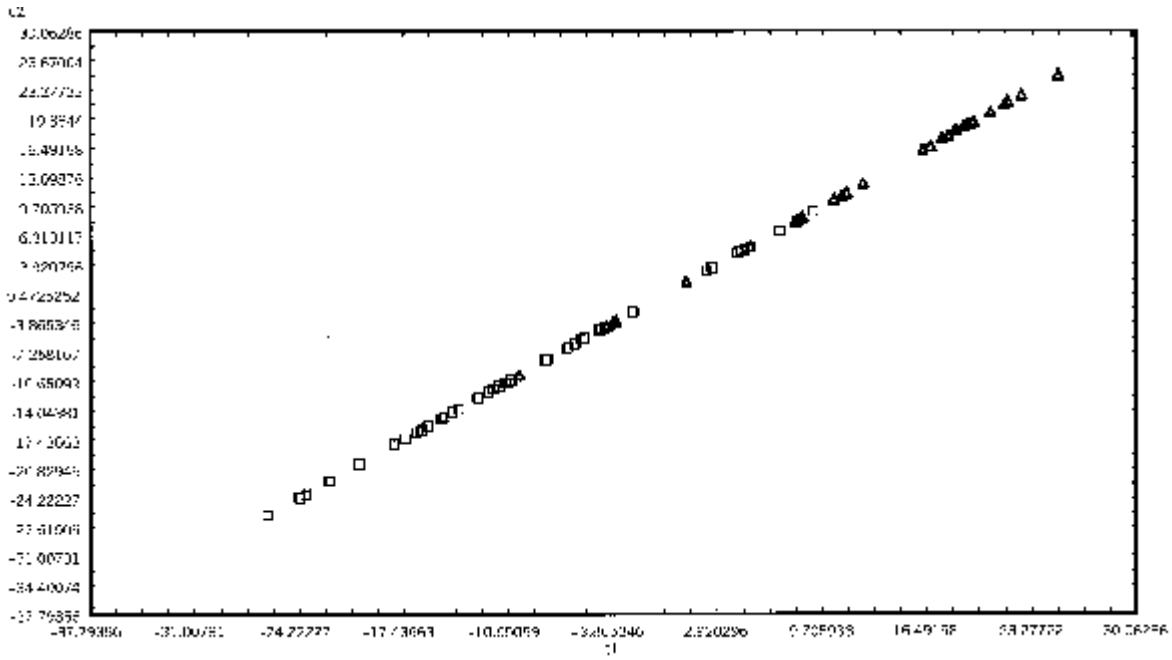


Figure 8 – A plot with the mysterious latent variables of ANN

I wrote to the Editor in Chief of the Journal, complaining about the poor quality of the work. He answered, more or less: “You have a forked tongue. I am not responsible, the responsibility is of the continental editors, which do not control the papers, of the referees, which evaluate a paper in five minutes, of the professors, which do not control the work of the students, and of the students, who are idiots”.

It is advisable that people using Chemometrics should work with:

true problems,

a clear understanding of the objectives,

the use of suitable Chemometrics techniques,

the use of optimal sampling procedures.

The above statements seems obvious, but these commandments have been violated an incredible number of times. By my own admission, I have worked a lot on the characterization of typical foods and I'm not proud of many of initial studies because I had violated in varying degrees the golden rules above. As a sinner, I'm lenient toward the sins of others. However, over time knowledge has increased and what may have been tolerated four decades ago is not acceptable today.

Some years ago I read a paper discussing extra virgin olive oils samples from both a European country (EC) and from a South America state (SA). All of the samples were collected in the same year. They were analyzed, on the same day by means of a powerful technique producing many hundreds of physical quantities. A classification method was applied to these EC and SA oils.

I explained the utility of this classification model to my students as follows:

Notoriously in a supermarket there are oil bottles without labels. When a buyer finds one of these bottles, the question immediately arises: "Is it from EC or from SA?"

A second remark regards the representativity of the samples. No information about the sampling design was given in the paper.

A third remark is that the oil changes with time. The age difference between EC and SA oils was six months, a large interval compared with the shelf life, 12-18 months.

A further important remark is about the technique. Classification methods work at the first level of pattern recognition.

Remember [14] that the levels are:

(1) Classification into one of a number of defined classes.

(2) Definition of a class model and of an acceptance-refusal rule, plus (eventually) level 1.

(3) Ability to relate the variables measured to one external property of continuous character.

(4) Ability to relate the variables measured to more external properties.

What is necessary in food chemistry (and in quality control) is the use of class modelling techniques. The bottles of oil in the supermarket always have a label. The true problem is control of the label, of the authenticity, the origin and the quality. The class modelling techniques, as SIMCA, UNEQ, Potential functions and CAMM, answer this objective.

When we define a category, the samples must represent all of the variability factors in the category. A valid class model can be built when the olive oil samples represent the different climate characteristics, soil, exposition, olive collection procedure, time of collection, storage time, pressing technique, time from pressing, oil storage conditions and year of production. Whilst this constitutes a very large and expensive collection of samples it is arguably better to join the efforts of many people to realize a really useful goal rather than publishing a lot of useless research papers.

Sampling is a crucial point of many, too may, research works in food science. I remember a paper where the Authors said "the samples were bought in the supermarket in front of the University". From this I have coined all types of insufficient sampling: "supermarket design".

On the positive side, Chemometrics searches for parsimonious models. A thrifty model has generally the advantages of larger stability and economy. Moreover, simple models can be understood easily.

On the negative side, too complex models and too difficult statistics can generate confusion and may lead to rejection of Chemometrics.

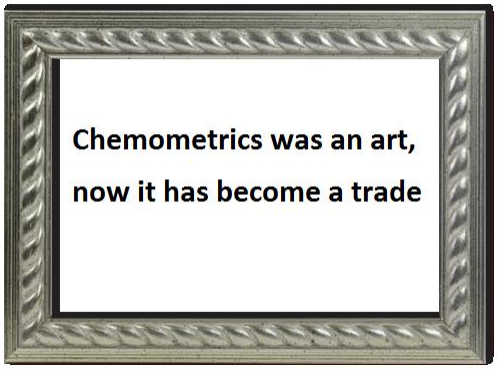
Chemometricians must remember that:

"Chemometrics must not be separated from chemistry, or even be allowed to become a separated branch of chemistry. As chemists, we must realize that we must continue to keep the power over our own theory, data analysis, model interpretation, and most important of all, our problem formulation. This means that we

borrow methods and tools from mathematics, statistics, etc. when and where we want and have the need to, but not because we feel obliged to. If we, wrongly, see the philosophy of statistics as stronger than that of chemistry, we will take the same sad and unfortunate route as biometrics, psychometrics,… which today are of little or no interest to biologists, psychologists,… This misfortune is the consequence of seeing mathematical and statistical "rigor" as more important that solving scientific problems. Of course, one should always have as much rigor as possible, but not "rigor mortis": chemical relevance comes first."

These teachings of Svante [15] after 20 years are still very much alive.

I add to the words of Svante a further consideration: enthusiasm is almost gone, and without it Chemometrics is intended to be covered by a tombstone



On the positive side, Chemometrics carefully validates the models.

During my first meeting with Bruce Kowalski he said:

"The force of Chemometrics is that our results are reliable, honest. We never overestimate the performances of a model. We validate."

Validation is a very important step in Chemometrics. It takes time and this can be tedious but careful validation is necessary and the time spent in validation is well rewarded.

Validation is not only for predictive ability. The double-cross validation of Svante Wold [16] should be used to detect the number of significant components.

On the negative side, validation is not satisfactory. When we use cross validation, it is necessary to use not only leave-one-out, but three, five, seven cancellation groups, many times with different object orders. Montecarlo validation with a lot of evaluation sets and objects can be very useful. The single test set (that we used when Chemometrics was in its infancy) is acceptable only when we have many objects and when both sets are very representative. Without these components the evaluation of the performance of the model is not reliable. The chemometricians must remember that the predictive ability is an experimental estimate and all of the estimates has a confidence interval. This can be obtained by the dispersion in the cancellation groups of cross validation or of Montecarlo validation. In the case of classification and class modelling there are suitable statistics [17] that compute the confidence interval

In that regard I have two examples.

The first is a joke: one of my students said to a companion “You had better results than me, but I created my test set by random extraction as you did”. The companion answered “I worked very hard. I created many test sets and finally I selected the set with the best prediction ability”.

The second example is more complex. Many years ago in a collaborative study, the participants decided to work with the same test set. They put the objects in the order of the response (moisture) then they selected for the test set one object every three. All the participants detected one Y-outlier in the training set. After the elimination of the outlier, with the new model they detected one Y outlier also in the test set (Figure 9). The conclusion was that the procedure gives good results, about 0.3 standard deviation of the prediction error. However, rarely and because of unknown causes, the prediction error is very large, ~4. Obviously such a foolish method can’t be accepted for the use in the analytical laboratory. Y-outliers are extremely dangerous.

When we performed cross validation on the same data without an external test set and without modification of the original order of the objects i.e. the order of the analysis we obtained the result in Figure 10. The two “outliers” were analyzed consecutively, and when the result of the analysis was interchanged, both of the outliers disappeared.

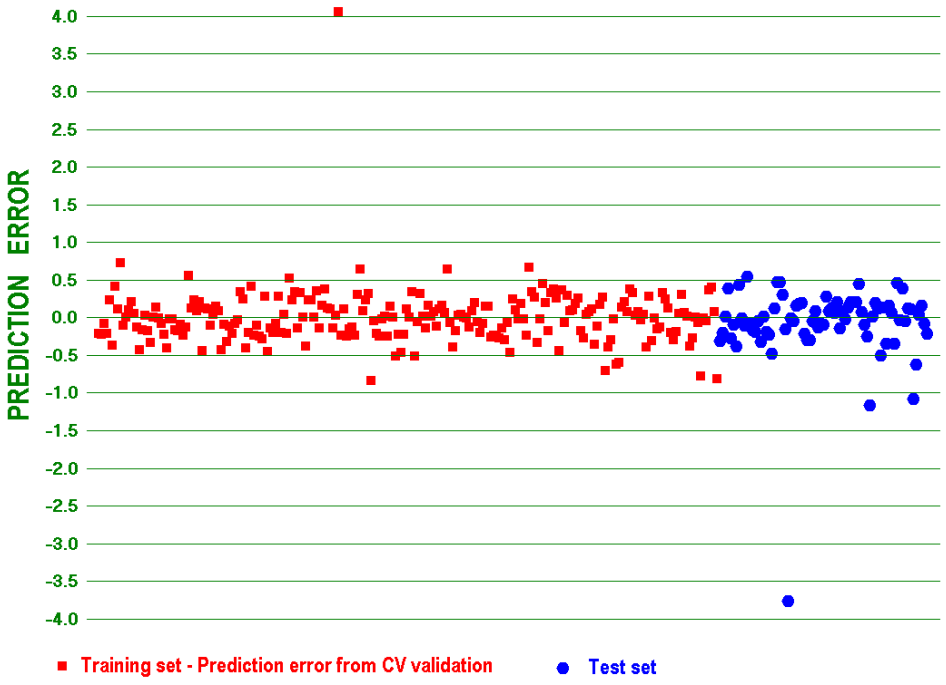


Figure 9 – Prediction error with the external test set

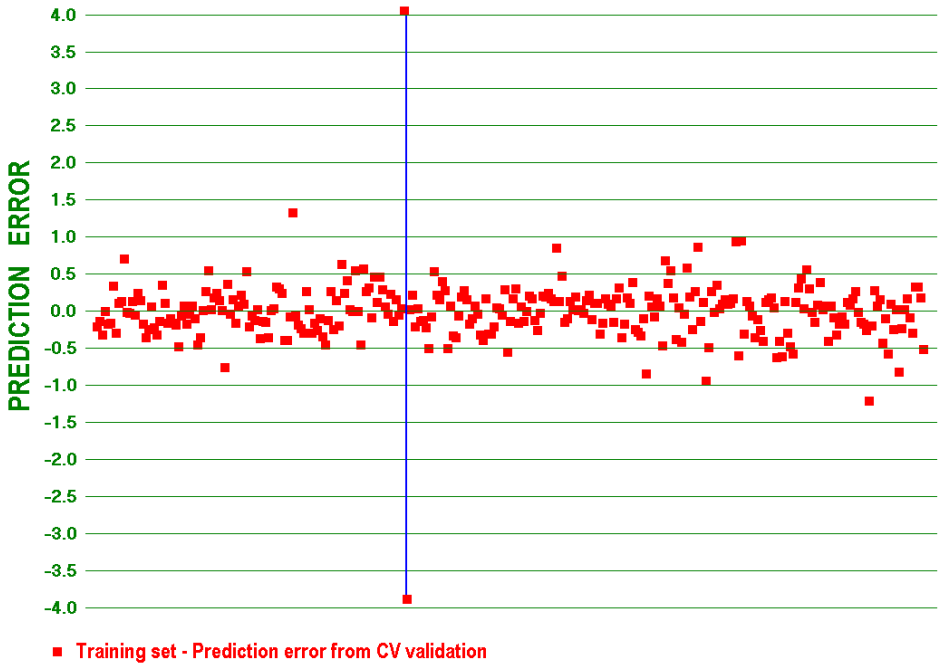


Figure 10 – Prediction error with cross validation

The example provides two important pieces of advice: firstly, be cautious with the unique test set; second, never lose contact with reality, the laboratory and the chemical significance of the data.

In these fifty years, the work of many chemometricians has produced an important improvement in the quality of research moreover the application of Chemometrics in real problems has expanded greatly and so I am optimistic for a positive future of good Chemometrics. However be warned, the spectra of dark Chemometrics has also made great progress and I hope that young chemometricians will be able to prevent its momentum!

Perspectives

Many years ago, during a talk with Luc Massart, I remarked something along the lines of:

“Chemometrics is now an important part of Analytical Chemistry”.

Luc corrected me:

“Chemometrics is Analytical Chemistry”.

Now I affirm “Chemometrics is Chemistry”.

It is impossible to work in Chemistry without previous information and it is impossible to obtain information from measured or computed data without data analysis. It is impossible now and it was impossible one century ago. The difference is that the amount of information increased dramatically, the structure of the information has become more complex, the retrieval of information has become easier and faster, the tools for handling the information has improved, all of this organized by Chemometrics.

Chemists work to solve problems.

This seems an evident statement however a lot of people work in chemistry with only one objective: to write an article and to be published in a journal with high impact factor.

There are a lot of variety of chemical problems. To solve a chemical problem it is necessary to have a chemist with a deep knowledge of the problem, it is necessary to be the “expert of the problem”. Then it is necessary to perform operations (e.g. synthesis) and always there are associated with these operations some measurements. So it is necessary to be an “expert of process”, and an “expert of instruments”. Operations and measurements are made upon samples. We need to be an “expert of sampling”. Operations and instruments produce data, so that we need to be an “expert of data analysis”, in other words: a Chemometrician.

It is very rare that a single person is, at the same time, an expert of problem, processes, instruments, sampling, and data analysis. So generally, to solve a true problem, we need a team. Here an expert must have a minimum knowledge of the other fields of expertise to make effective the collaboration work.

The data analysis expert must be a Chemometrician with a deep knowledge of chemistry, instruments, sampling, and processes. The role of the Chemometrician in the team is, in my opinion very appealing. Obviously this is a very subjective opinion. The pleasure obtained by the knowledge of almost all of the parts and steps of the problem, the delight experienced from the useful information flourishing from the chaos of the rough information, the set of sublime sensations experienced by the creation of a new tool for a new problem (including modest or modified tools) gives our work something incomparable. These are the reasons I became and remain a Chemometrician. These are the motivations of many other Chemometricians.

All of the experts in the chemical problems “team” can find many sources of pleasure in their work. I was born as an electroanalytical chemist and for many years the possibility of introducing improvements to expedite useful results was a source of satisfaction. However when I become a Chemometrician it was better. During the last fifty years, the expansion of chemical problems, of instruments and of computing power has been astonishing. When, fifty years ago, some prophets forecasted these developments, they were mocked but their daredevil predictions were often less potent than the future reality. In 1963 I worked on an IBM mainframe with 32 Kbytes memory with punched input and output cards. At a time a real wonder but now my personal computer does in two minutes the work the IBM would perform in one day. In around 1970, my team was reduced to poverty by purchasing an expensive Calcomp plotter. A wonderful instrument. Today a 50 € printer produces better plots in the time it took to change the pen on the Calcomp plotter.

I’m short sighted. However I think that some of today’s Chemometricians applications will have interesting developments: image analysis, metabolomics and synthesis are among today’s frontiers. We study for example images to evaluate a product, to make a diagnosis, all very interesting and very useful. The data matrix is very large with millions of numbers but we have another dimension, time. This means billions of numbers. Over time we can study the evolution, and predict. Today we are in the same position as the

weather forecast fifty years ago. We are watching from our window. It is a useful place to observe but we can do more.

I worked for many years on the data analysis of food. Even in this field, almost always, data describe a specific moment in the life cycle of the product which is in continuous evolution. To study the evolution from the marketing to the expiration date it is necessary to repeat the measurements many times what is heavy and expensive but gives validity to the research.

I'm short sighted, but I see an interesting future for young chemometricians. If I had to start my life over, I would be a Chemometrician, again.

References

[1] B.R Kowalski., P.C. Jurs, T.L Isenhour, C.N., Reilley, Anal Chem., 41, 695 (1969)

[2] P.C. Jurs, B.R. Kowalski, T.L. Isenhour, C.N. Reilley, Anal Chem., 41, 690 (1969)

[3] H.Martens, H.H.Russwurm Eds., "Food research and data analysis", Applied Science Publ., Barking, 1983

[4] B.R.Kowalski Ed., "Chemometrics: Mathematics and Statistics in Chemistry", NATO ASI Series, Ser.C, Vol 138, 439-466, Reidel Publ.Co., Dordrecht, 1984.

[5] M.P. Derde, D.L. Massart. "UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution". Anal. Chim. Acta, 184, 33-51 (1986).

[6] M.Forina, S.Lanteri, L.Sarabia, J. Chemometrics, 9, 69 (1995)

[7] M.Forina, C.Armanino, R.Leardi, G.Drava, J.Chemometrics, 5, 435 (1991)

[8] C.Pizarro Millan, M.Forina, C.Casolino, R.Leardi, Chemometrics and Intelligent Laboratory Systems, 40, 33 (1998)

[9] M. Forina, S. Lanteri, M. Casale, M.C. Cerrato Oliveros Chemometrics and Intelligent Laboratory Systems, 87, 262 (2007)

[10] W.S. Robinson, American Antiquity, 16, 293 (1951).

[11] M.Forina, C.Casolino, C.Pizarro Millan, J.Chemometrics, 13, 165 (1999).

[12] M. Forina, M. Casale, P. Oliveri, S. Lanteri, Chemometrics and Intelligent Laboratory Systems 96, 239 (2009)

[13] M. Forina, Fondamenta per la chimica analitica, (Italian), ISBN 9788890406461, Edited and distributed free from SISNIR through its website

[14] C. Albano, W.Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, "Four levels of pattern recognition", Anal. Chim. Acta, 103, 429–443 (1978)

[15] S. Wold, "Chemometrics; what do we mean with it, and what do we want from it?", Chemometrics and Intelligent Laboratory Systems, 30 (1995) 109-115

[16] S. Wold, "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models", Technometrics, 20, (1978) 397-405

[17] J.K. Martin, D.S. Hirschberg, Technical Report No. 96-22, 594,University of California, Irvine, 1996.