Chemometrics; what do we mean with it, and what do we want from it?

Svante Wold

Institute of Chemistry, Umeå University, S-901 87 Umeå, Sweden

Received 4 January 1995; accepted 10 May 1995

Abstract

What we mean with chemometrics is closely connected to what we want from it, our idealized view of chemistry, data, statistics, models, and what not. Personally, I have a fairly clear opinion of what I mean with, and what I want chemometrics to be. Of course, this view is based on my experience in academic and industrial chemistry, and I shall here try to explain these views, and the chemical and practical reasons (rationalizations) thereof. These views may seem somewhat provocative, but this is not the objective. I just want to make us reflect on what we, as chemometricians, are doing.

My views will be focussed on the role of chemometrics in chemistry, on how I see statistics as partly consistent, and partly inconsistent with the framework of chemistry, and where we should aim for the future (of chemometrics).

Keywords: Chemometrics; Design; Multivariate analysis; Statistics

1. What do we mean by ‘chemometrics’ (and what do we want it to be)?

In 1974 (20 years ago) as a young aspiring docent (assistant professor) I wrote the following: ‘The art of extracting chemically relevant information from data produced in chemical experiments is given the name of ‘chemometrics’ in analogy with biometrics, econometrics, etc. Chemometrics, like other ‘metrics’, is heavily dependent on the use of different kinds of mathematical models (high information models, ad hoc models, and analogy models). This task demands knowledge of statistics, numerical analysis, operation analysis, etc., and in all, applied mathematics. However, as in all applied branches of science, the difficult and interesting problems are defined by the applications; in chemometrics the main issue is to structure the chemical problem to a form that can be expressed as a mathematical relation. The connected mathematical problems are rather simple. (Today, 1994, I would like to add: ‘as the statistical problems usually are’). Therefore, chemometrics must not be separated from chemistry, or even be allowed to become a separate branch of chemistry; it must remain an integral part of all arcas of chemistry.’

However, in 1994 I cannot say it much better, and a reasonable definition of chemometrics remains as: ‘How to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data’.
2. So, how do we translate this into something practical and workable?

With the 'definitions' above we end up with semantic questions about chemical, relevant, information, representation, etc. However, we can still make some fairly strong statements about chemical information (CI). The only reasonable way to extract and represent this CI is in terms of models. The relationships between, on the one hand, chemical macroscopic variables such as concentrations, temperature, pressure, viscosity, and, on the other hand, microscopic (molecular) variables such as molecular surface area, molecular volume, substituent size, charge distribution, etc., are from necessity approximate and valid only within certain intervals of the variables. It is, for instance, obvious that the same relationship between water accessible surface area and solubility cannot hold for, say, both halogenated hydrocarbons and hexose monosaccharides. Hence, these relationships are best thought about and treated as approximate models.

A chemical model (M) relating experimentally determined variables, X, to each other, in addition has a statistical model (E) associated with it. This model, E, describes the variability, the noise, of the data around the chemical model. This separation of 'chemistry' and 'noise' is an old invention of statistics, and helps us enormously when we want to conclude anything from measured data

\[ X = M + E; \text{ Data} = \text{Chemical Model} + \text{Noise}\]

Although this model concept has been used in statistics for a long time, it still is not well accepted in chemistry, except in its analytically oriented branches. Still, 70 years after Bohr and Heisenberg, we prefer to think in terms of natural 'laws' and the chemical 'truth'. This causes a communication problem between chemists and chemometricians which, however, seems unavoidable.

As soon as we analyze any chemical data of sufficient amounts and accuracy we find systematic deviations between the data and the chemical 'model' or 'law'. Chemical theory then takes recourse to idealized states such as infinite dilution, gas phase acidity, etc. This is impractical both for interpretation since the idealized states usually are experimentally inaccessible, and for practical use since we are interested in predicting what will happen at finite concentrations in imperfect solutions.

The model concept provides a more realistic framework for chemical data analysis than the often misleadingly exact looking relationships derived from first principles. In particular multivariate 'empirical' modelling by PCA and PLS, often shows unexpected patterns because the joint effect of all variables is taken into account, in contrast to traditional chemical and physical relationships which usually consider just one or a very few variables at the same time.

However, as chemometricians we must remember that chemistry is our center and our base, and that chemometrical approaches without acceptance in chemistry are of little value. Here great educational efforts on all sides are necessary. Also, lots of interesting research is needed to make the 'fundamental' models used in all branches of chemistry consistent with the presence of noise, co-linearities, and model errors.

3. Chemistry is a study of complicated systems

Any interesting chemical system or process is too complicated to be understood 100% (or even 99%) by theory. Hence, the experimental approach to achieving information is necessary in chemistry, and even more so in the interfaces chemistry/biology, chemistry/technology, and chemistry/the real world. This does not exclude theory and experience; they tell us what to measure and how and where to sample, i.e., they take us as far as a certain domain in an abstract multidimensional space. The co-ordinates of this space are partly known, and partly unknown, because we do not fully know how to adequately represent a chemical problem. For instance, we cannot translate structural variation among a set of molecules to relevant coordinates that well correlate with the variation in measured properties such as catalytic ability, taste, or antibacterial activity, routinely.

Importantly, the domain, the multidimensional volume, containing the 'best' solution to a given objective has a nonzero volume. This necessitates experiments to find the best point, the best set of conditions inside this domain. Also, experiments are needed if we want to study, interpret, and understand...
the variation of properties in this volume. And to connect the experimental results (data) we need empirical models, because theory takes us only to the boundaries of this domain. Of course these empirical models need to be consistent with the theory, at least on the boundaries of the domain.

With simple principles of experimental design, experimentation can be made to investigate this domain magnitudes more efficiently and more 'informationally' than is presently done with the traditional 'one factor at a time' approach. For chemometrics, teaching and further developing these principles in chemical investigations is a most important task.

In the next decades of chemometrics, there is a great challenge to make the principles of experimental design applicable in all areas and all types of problems in chemistry. The greatest challenge may be to make this understandable and palatable to every active chemist.

4. The study of complicated systems is by necessity multivariate

Another dominating (and closely related) issue is the rich, often not recognized, information in multivariate data; multidimensional data measured on a set of similar samples, cases, objects, process points. The simple principles of models, multidimensional spaces, and projections provide a strategy for utilizing this richness of information.

The early interests of chemometrics – pattern recognition, classification, linear and non-linear mapping – are still relevant. And today we start to understand the properties of chemical data, and how to modify the chemometric methods to correspond better to these properties.

The next generation of chemometrics methodology will include improved PLS and other modelling approaches with better predictive power, and more simple interpretation.

From a philosophical viewpoint, it is interesting to see that much of the chemometrics methodology is based on the ‘indirect’ approach of investigation. Experimental design can be seen as an indirect approach to the study of the joint effects of many factors. Multivariate analysis is the indirect observation of intrinsic, ‘latent’, variables.

With this in mind, it is strange that mainstream chemistry has such difficulties to accept chemometrics and the indirect approach to experimentation and data analysis. After all, the whole of chemistry can be seen as the indirect study of ‘latent concepts’ such as molecules, atoms, electrons, and molecular reactions. None of us has seen a molecule more clearly than we have seen a latent variable.

5. Variability, noise, statistics

Any measured data has noise, uncertainties, variability, and interactions, non-linearities, and unrecognized features (model deficiencies). This includes anything measured in chemistry. And mainstream chemistry has still great difficulties in recognizing these basic facts, as have, strangely, many so-called chemometricians. And the art of drawing conclusions – extracting information – from noisy data we call statistics, which we therefore need, regardless if we like it or not.

6. The role of statistics:

Whatever we measure in chemistry (as well as in physics, biology, geology,...) is imprecise. Any data, any measurement, has some variability or noise. I prefer these terms before ‘error’ which has a negative connotation, and hence encourage us to forget it or repress it.

Statistics is the art of drawing conclusions from data, and making decisions, in the presence of variability. However, in my view, statistics does not yet fully live up to these ambitions, because it usually does not recognize the importance of model errors. Of course there is good statistical research done on how to recognize and deal with model errors, but this has not yet reached neither standard textbooks nor standard software packages that chemists are exposed to.

The deviation between a measured value and our ‘model’ is due to at least three sources. Firstly, our lack of complete control of experimental conditions
(temperature, pH, concentrations of this and that, etc.) makes it impossible to exactly maintain the same conditions the second time we measure something or the second time we make an experiment. We may be determined to set temperature to 25°C, pH to 4, and the reactant concentration to 0.1 M. But the partial lack of control makes the temperature to be 24.96°C and 25.02°C in the first and second observations, pH to be 4.02 and 3.97, etc. The second time the measurement will be made under slightly different conditions, and hence give a slightly different value.

Secondly, the measurement instrument – a scale, a spectrometer, a chromatograph – will not be exactly stable, and will produce a slightly different value the second time we measure 'the same thing'. This is because the instrument is influenced by lots of secondary factors; the barometric pressure, ambient humidity and temperature, the phase of the moon, the voltage in the electrical outlet, and what not. Together these factors will cause a significant variability; the measurement variability (error). This 'measurement variability (error)' is well recognized both in chemistry and statistics as a source of variability, and often seems to be the only 'legitimate' source of variability in measured data.

In the Taguchi school, these two sources of variability are called 'noise', and 'environmental' factors, respectively.

Thirdly, the model errors. These are due to the inescapable simplifications and approximations inherent in our scientific models such as reaction mechanisms, Lambert–Beer law, Debye–Huckel law, etc. We use our scientific models to specify what we expect from our measurements, and any deviations we interpret as 'errors of measurement', not as a deficiency of the model. For instance, constant concentration, constant spectrum, etc. are models, based on which we estimate average concentration over six samples, the average spectrum over 14000 scans, etc. But, of course, this idea of 'constancy' is a model, an idealization, which is not altogether achievable.

Hence we can safely conclude that any chemical data we put our hands on are 'contaminated' with noise. Much of this noise is random (whatever we mean with that) but parts of it (the model 'errors') is systematic, non-random. The corollary is that we need statistics for the analysis of chemical data, because statistics is the art of dealing with noise.

7. Some observations

After many analyses of large data sets from various projects and processes, always with the great hope that the 'secret' will be revealed by this analysis, I have realized that the typical results of such analysis is that there is very little there. The same experience has been experienced by George Box, Cuthbert Daniels, and everybody else analyzing real data. This experience can be stated as: 'The low quality of a data set is not realized until it is properly analyzed'.

The insight that chemical (and other) data are noisy, with the consequences for the 'absence of truth', is still rather rare in chemistry. Those who try to instill this insight will hence run the risk to look as anarchists who are negative to all 'law and order' in chemistry. There is a risk that chemometricians, like statisticians, will be seen as petty negativists who insist on confidence intervals around all resulting parameters, who question all experimental results and their interpretation, etc. Diplomacy, patience, and remembering our chemical heritage helps.

8. What is good and bad (for chemistry) with statistics

The modelling of noise in terms of distributions (normal, log-normal, t, F, Barndorff–Nielsen, etc.), is a great contribution to science and thereby chemistry. This has lead to the concepts of confidence intervals, t-tests, F-tests, X²-tests, outlier tests, etc., all very useful in chemical practice.

The model concept is a second great contribution of statistics. With models we see data as an imperfect model plus noise, which is very fruitful both for the design of experiments and measurements, and for the interpretation of measured data. And model fitting to data by means of least squares has given science a well working formalism for the analysis of chemical, physical, biological, geological, and other measured data. The generalizations to weighted least squares, non-linear least squares fitting, etc., makes the least squares methodology generally applicable.

However, other concepts central to modern statistics do not, in my view, apply equally well in chemistry and chemometrics. 'Populations' are relevant to
genetics, but usually not to chemistry. We are not interested in the average boiling point of the ‘population’ of organic esters, but rather in the boiling point of each individual ester as a function of its structure. Perhaps, we can discuss the variability of an analytical chemical determination in terms of a ‘population’, but I don’t feel that this conceptual model really improves our understanding of variability. In genetics there is a real variability between the individuals in a concrete existing population of, say, cats. In chemistry we do not have a concrete population of determined values of, say, a $pK_a$, until after some attempts to make this determination. And we can not remeasure the same individuals in the same way as we can remeasure a property of our cats. Hence, there is no real population of $pK_a$ determinations with an independent existence that we can sample, measure, remeasure, etc. Here the population concept is at best an abstract model of a real population, in my view lacking so many properties of a real population that it is not very useful in chemistry. Indeed, my experience with teaching variability to chemists is that we do better without it.

Similarly, seeing measured multivariate data as coming from a (multivariate) distribution is not consistent with the philosophy of chemistry. We are not willing to think of chemical properties, such as melting points, viscosities, or enzymatic reaction rates, as drawn from multivariate distributions. Rather, we divide data into a systematic (deterministic) part which, in principle, can be precisely known, plus a random part, the noise, the distribution of which is often less well understood. This is indeed the standard statistical way to model few-variadic data; in chemometrics we make the same explicit separation also with multivariate data, as is done in psychometrics.

In essence, the experimental control of chemical systems makes observations made on these systems be partly non-random. Philosophically, it is also preferable to see data as a chemical model + noise instead of as a single multivariate distribution, because the former emphasizes the chemistry. And it is difficult to include experimental designed data in the distribution framework; their $X$-part is definitely not random even if they are multivariate.

Strangely, in regression modelling the $X$-part (predictor variables) is normally considered ‘exact’ in statistics, while in chemometrics we normally consider them as being noisy. This makes PLS- and PC-regression preferable to classical multiple regression in the analysis of data sets with many and collinear variables. Also, in traditional statistics the $X$-variables in regression are considered as mathematically ‘independent’ (i.e., the $K \times N$ dimensional $X$-matrix is assumed to have rank $K$). In chemometrics we prefer to estimate this rank from the data and usually find it appreciably smaller than $K$. Here a semantic problem adds to the confusion. It seems that ‘independent variables’ mean different things in statistics (orthogonal variables) and in mathematics (variables with a correlation less than 1.0).

Thus, in PCA, PLS, etc., we prefer to see scores ($T$) and loadings and weights ($P, W, C$, etc.) as systematic, deterministic, not derived from a ‘population’ of random samples.

$$X = TP' + E$$
$$Y = TC' + F$$

These parameters describe our chemistry, and are, in principle, interpretable in terms of chemical concepts. Only the residuals ($E$ and $F$) are partly random, but usually also partly systematic, making their distribution difficult to specify, and making least squares the preferred estimation method. Certainly, any estimation of $T$ and $P$ from real noisy data will make the estimated values contaminated with some noise, but the ‘real’ $T$ and $P$’s we think about as deterministic.

Of course there are cases when we can see also $X$ and $Y$ as (partly) random, for instance in multivariate calibration of protein and water in grain, when grain samples are drawn more or less randomly to form a calibration set. Continuing to see $X$ and $Y$ as mainly non-random also in this case, however, does not seem to matter much for the data analysis and interpretation of the results, and most chemists, including myself, seem to prefer seeing any measured data as mainly deterministic with only a small random part to which statistics applies.

It seems that the whole classical statistical multivariate analysis (except factor analysis) is based on the assumption that multivariate data can be considered as drawn from a multivariate distribution. Hence much of multivariate analysis has to be reinterpreted and even to some extent redeveloped for its appropriate use in chemistry.
One of the cornerstones of modern statistics is the concept of maximum likelihood (ML), and ML estimation has completely overtaken least squares estimation. ML estimation is based on the idea of model parameters that give the data (or residuals) the distribution that best corresponds to the 'true' distribution. ML estimation was developed as a reaction to misuse of unweighted least squares estimation in the 1920s and 1930s. Thus ML explicitly accounts for the distribution of the residuals, particularly when they have widely different variance in different areas of the experimental domain. With modern diagnostics of least squares (LS) for finding outliers, grouped observations, and non constant variance of residuals, LS estimation works as ML estimation. In addition, LS estimation is consistent with non-random residuals due to model inadequacies, while ML estimation is not. And from the practical point of view, LS estimation computes usually magnitudes faster, and is much easier to program.

In this context, the concept of unbiased estimators is equally dubious. These estimators have the theoretical means of parameter estimates equal to the 'true' parameter values. But what happens, when there is no true value, and the model underlying the estimator, is wrong. 'Biased' estimators such as ridge regression, PCR, and PLS, in addition, have estimates with smaller variance, i.e. better precision and accuracy, than the 'unbiased' estimator for a given N.

To summarize; we know that all models are wrong. How can we then motivate ML estimation, and what is meant by unbiased models? It seems that the chemical and chemometric approach to use least squares, weighted when motivated, data transformed when needed to make residuals more symmetrically distributed, etc., is a more practical approach than that of modern statistics, at least for messy and complicated problems.

9. A modest forecast

In the near future we can see a number of exciting application areas of chemometrics open up even more. Analytical chemistry has the modelling, the improvement of separation (chromatography, electrophoresis, etc.), and instrumental methods. The modelling and optimization of syntheses, structure property relationships, and composition property relationships, and the interpretation of spectra, are of great principal interest in organic and inorganic chemistry, biochemistry and, perhaps, even in physical chemistry. Engineering and environmental chemistry have, in addition, process monitoring, modelling, and control, as central areas of interest.

A 'chemometric' view of physical chemistry would encourage the evaluation of physical chemistry 'properties' in terms of their informational content, i.e., how they help to distinguish classes of reactions, or compounds, or other entities of chemical interest.

However, a successful future for chemometrics depends, on a continued problem orientation, with the most important task seen as solving chemical information problems in an efficient and appropriate way. Certainly some method development is needed, but this should be motivated by solving real chemical problems, not by mathematical and statistical sophistication. To quote Einstein: 'Make everything as simple as possible – but not simpler'.

10. Conclusions

Chemometrics is the branch of chemistry concerned with the analysis of chemical data (extracting information from data) and ensuring that experimental data contain maximum information (the design of experiments). The future of chemometrics is bright. On average, the present state of data analysis and experimental design in both academic and industrial chemistry is, medieval, and therefore lots of chemometrics are indeed needed.

However, we must be careful that we do not lose chemistry in our zeal to reform it and make it more quantitative and realistic by taking noise and uncertainty into account. If we let statistics become too influential, its present fashions such as maximum likelihood, multivariate distributions, and other esoteric concepts, might convert chemometrics into a useless appendix of mathematical theory.

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, statis-
tics, physics, 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, and much of the technometrics, which today are of little or no interest to biologists, psychologists, engineers, etc. This misfortune is the consequence of seeing mathematical and statistical 'rigor' as more important than solving scientific problems. Of course, one should always have as much rigor as possible, but no 'rigor mortis'; chemical relevance comes first. In essence, we must remain chemists and adapt statistics to chemistry instead of vice versa. And chemometrics must continue to be motivated by chemical problems solving, not by method development.