CHEMO in chemometrics

Paul Geladi

Department of Organic Chemistry, Umeå University, Sweden.

This is a loose collection of thoughts about the "chemo" in chemometrics. The reader should not try to find too much structure in the text. The topic is much too wide to be treated in full in such a limited space. So there is no promise of an exhaustive treatment, but it is hoped that the ideas expressed here may serve as a source of inspiration for readers.

To find out what the chemo in chemometrics means, a look at the definition is in order. The definition used by NAMICS (the North American chapter of the International Chemometrics Society) is:

Chemometrics is the chemical discipline that uses mathematical, statistical and other methods employing formal logic

1. *To design or select optimal measurement procedures and experiments*
2. *To provide maximum relevant chemical information by analysing chemical data*

This definition is found in a book by Massart *et al.*[1] and in the NAMICS home page.

I have a problem with the term "formal logic". Logic is not a property of nature. It is a human invention that works locally in closed environments. Except for the likes of officer Spock and Sherlock Holmes, logic should have a very low status.

Something that is missing in the definition of chemometrics is computer science. I have mentioned this on a few earlier occasions.[2–4] Chemometrics rarely uses analytical solutions or simple calculations. Most solutions are iterative and all calculations are based on fast computation and efficient database management. Chemometricians are becoming more and more dependent on this. Also, visualisation in 3-D and colour is playing an increasingly important role. The role of communication in all aspects of life is, of course, also reflected in chemometrics. There are different names for this phenomenon, but "the internet" describes all these aspects fairly well. Other parts of chemometrics that fall mainly under computer science are neural networks and genetic algorithms. In Reference 5 our dependence on computers is mentioned.

The chemo part of chemometrics is not really well-defined. A possible extreme interpretation of the definition is that statisticians and mathematicians (computer scientists too) can just receive data on a diskette and do superior work. This is, however, not the case: fortunately for us. There are a

number of situations where the chemo comes in strongly. In the chemometrics literature this is not always emphasised. Impressive equations seem to be more attractive than a chemical interpretation and this gives the false idea that any mathematician or statistician could have done a better job. They could have done a better job with the equations, but not with solving the chemical problem. By the way, statisticians, mathematicians and computer scientists do excellent work. The fact that this text is about the chemo in chemometrics should not be seen as criticism of any other branch of science.

A classical definition of the chemo in chemometrics is that it refers only to analytical chemistry. This definition is incomplete, but chemometrics seems to have made its more spectacular and commercial advances in analytical chemistry. There is no harm in looking to other fields of chemistry to find out what more can be done. It would also increase the status of chemometrics.

The rest of this text contains some loose examples in which the chemo is superior to mathematical and statistical interpretations:

- rank determination and precision in regression analysis
- situations with outliers
- situations with missing data
- factor rotation and non-uniqueness in three-way factor analysis, the role of scaling.

For rank determination in regression analysis, mostly for partial least squares (PLS) and principal component regression (PCR) models, there are a number of criteria and tricks for separating the meaningful rank from the noisy part of the data. Cross-validation is very popular and extremely overrated. There is this false belief that since there is a number that can be calculated and that this number is "best", this must be the truth. One of the reasons for this misbelief is that it is easy to write software that gives this number automatically. All other methods of rank determination require some thinking. This does not mean that cross-validation is bad, just that it should not be used as the only method. So what is the chemical interpretation that should supplement cross-validation? Spectral interpretation of loadings is meaningful. Good examples can be found in Martens and Næs.[6] What is the chemical interpretation of the many PCR or PLS components that are often needed in near infrared spectroscopy calibration? Here there is no need to look for a mathematical or statistical solution. The solution should come from a better understanding of the chemistry and spectroscopy and this will require a lot of future work.

More about the chemo in calibration follows. When reading calibration literature it sometimes looks as if everybody is in a contest to find the lowest prediction errors. A few remarks need to be made here:

- No matter what You publish, somebody will modify the technique and find lower prediction errors.
- It is not meaningful to find prediction errors that are lower than natural variability.
- Most published improvements are not statistically significant.

The trick here is to find a good expression of natural variability. Any prediction error that comes close or falls below that is good enough. This is chemical common sense. There is also the fact that prediction error may not be the only thing looked for. Interpretation is equally important. This becomes a duality. A model that gives good chemical/physical interpretations may be different and have a different prediction error than one that gives good predictions.

A simple trick that has not been used very often in regression is that the error variance of the calibration and test samples should be the same. It is easy to construct an F-test for this. Even better, would be to compare the histograms of calibration residual and test residual with a χ^2-test, but this requires quite a lot of calibration and test samples. See Figure 1 for a schematic presentation.

For missing data, there is no unique one-size-fits-all solution. There are different situations with patterned and less patterned missing data. Missing data are too often treated as if they were randomly distributed over the data array. Many statistical and mathematical tricks are useful but, in some situations, common sense and a chemical interpretation are needed. One should be aware of the patterns and their sources. These sources may be related to the underlying chemistry.

The same is true for the special kind of missing data where measurements end up below the detection limit. This may seem trivial, but a number of factor analysis and other algorithms are very sensitive to the way in which below-detection-limit data are reported. Especially with non-linear methods, the choice of representation of below-detection-limit data is impor-

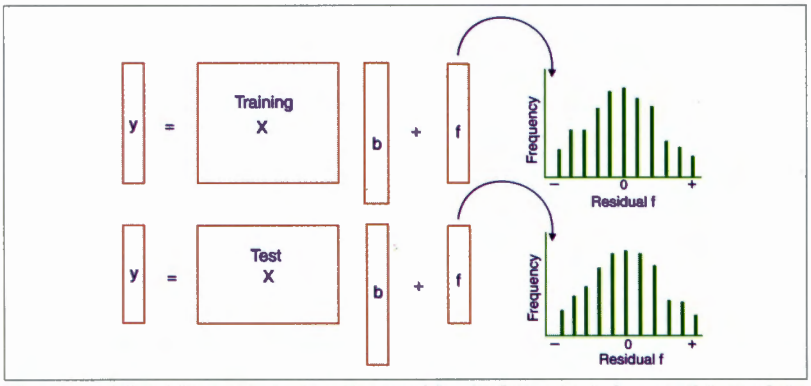

Figure 1. The histograms for the residuals from training and test data should look very similar.

tant. They should probably be distributed randomly between zero and the valid detection limit. Chemical common sense in interpreting detection limits is better than using standard recipes.

Outliers are also phenomena that should not be treated by mathematics or statistics. There are hundreds of statistical recipes for detecting outliers,[7] but unless there is a chemical interpretation, a mechanically found outlier does not make sense.

One of the fields where the "chemo" comes in very strongly is environmental chemistry. I was lucky to be on a sabbatical stay at Clarkson University, Potsdam, New York (Jan–June 1997). The work that they do there is based on environmental measurements,[8–11] often used with two-way and three-way factor analysis (see Figure 2). Here we have something interesting: the data and results should be interpreted as mass and ion balances. We also meet another duality (or maybe a Heisenberg uncertainty principle). No scaling of the data is correct. For ion and mass balances, everything should be expressed in molarity or normality. Any scaling that takes away normality does not give results leading to a chemical interpretation. So now we can talk about a number of different scalings:

- Unit-free (scaled by the standard deviation)

This scaling is "equal opportunity". Without this scaling trace elements/compounds always end up in the residual.

- Scaled in concentration (ppm, ppb, g L^{-1} etc)

This is how most results are expressed, and probably the least useful scaling of all. This is also how environmental regulations are expressed. Moles are not popular with legislators.

- Scaled in % or fraction (weight fraction, volume fraction, mole fraction?)

For some problems with mixtures this would be correct. Phase diagrams in physical chemistry are often expressed like this.

- scaled in molarity/normality (but unless relatively dilute solutions in a well-behaved neutral solvent are used, molality would be better).

This is the only correct scaling for interpretations of ion and mass balances.

The scalings influence the analysis. They also influence the interpretation of parameters after the analysis. This is an extremely difficult matter, especially when many-way data are considered. (see Figure 2)

Another interesting observation is that orthogonality is a bad thing. Concentrations and source contributions are all above zero. Introducing orthogonality forces certain values to negative and that makes no chemical

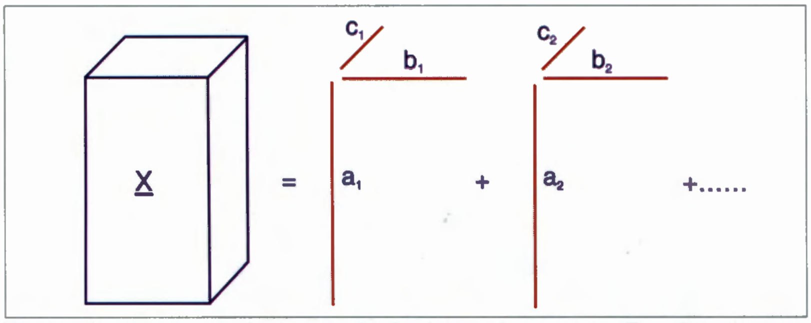

Figure 2. Three-way factor analysis, here shown as a Parafac decomposition, allows many scaling types.

sense. A duality develops again. Methods with orthogonal factors are easily and quickly calculated. They are quite unique too. The real models with non-negative factors are difficult to calculate, non-robust, non-unique and slow. The secret of getting anything at all is a chemical interpretation of the results.

Figure 3 shows an impossible chemical reaction. This is immediately clear. Multivariate analysis with bad scaling and a mathematical or statistical interpretation often contains reactions like these implicitly. The complexity of the multivariate model effectively hides what is really going on. Chemical interpretation is absolutely necessary in order to avoid blunders of this type.

Chemometrics should never be an excuse for bad sampling or sloppy analyses. There have been occasions where information was actually extracted from bad data, but by doing the sampling correctly and performing the best analysis possible, one gets more robust and informative models. This is also part of the chemo in chemometrics.

An impossible chemical reaction

$$\underset{1\text{ gram} + 1\text{ gram}}{Na + Cl} \longrightarrow \underset{2\text{ grams}}{NaCl}$$

Figure 3.

Acknowledgements

I would like to thank Barry Lavine of Clarkson University, Potsdam, NY, for some interesting discussions on the nature and definition of chemometrics. Phil Hopke of Clarkson University, Potsdam, NY and Pentti Paatero of the University of Helsinki are acknowledged for discussions on three-way factor analysis of environmental data.

References

1. D. Massart, B. Vandeginste, S. Deming, Y. Michotte and L. Kaufman, *Chemometrics: a textbook.* Elsevier, Amsterdam (1998).
2. P. Geladi, *Analysis Europa*, **April,** 34 (1995).
3. P. Geladi and E. Dåbakk, *J. Near Infrared Spectrosc.* **3,** 119 (1995).
4. P. Geladi, *Chemometrics in Belgium-Newsletter* 2 (1995).
5. P. Geladi and A. Smilde, *J. Chemometrics* **9,** 1 (1995).
6. H. Martens and T. Næs, *Multivariate Calibration.* John Wiley, Chichester (1989).
7. V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. John Wiley, Chichester (1994).
8. Y. Zeng and P. Hopke, *Atmospheric Environment* **26A,** 1701 (1992).
9. Y. Zeng and P. Hopke, *Chemometrics and Intelligent Laboratory Systems* **7,** 237 (1990).
10. P. Paatero and U. Tapper, *Chemometrics and Intelligent Laboratory Systems* **18,** 183 (1993).
11. P. Paatero and U. Tapper, *Environmetrics* **5,** 111 (1994).