Chemometrics: Fads and Fallacies

Richard G. Brereton
School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, U.K.

Abstract

Chemometrics as an identified methodology has been around for over forty years. During the last couple of decades, there has been a large growth in available software, meaning that many people using chemometrics methods no longer have fundamental statistical or computational expertise, and often do not understand the basic approaches they are using.

Discussion

In the biological literature, statistics is routinely used often misunderstood. Writing in Nature, David Vaux [1] states "*...it is still common to find papers in most biology journals contain basic statistical errors. In my opinion, the fact that these scientifically sloppy papers continue to be published means that the authors, reviewers and editors cannot comprehend the statistics, that they have not read the paper carefully, or both. Why does this happen? Most cell and molecular biologists are taught some statistics during their high school or undergraduate years, but the principles seem to be forgotten somewhere between graduation and starting in the lab. Often, the type of statistics they learnt is not relevant to the kinds of experiment they are now doing. And, once in the lab, people generally just do what everyone else does, without always understanding why*". Even (or especially) high impact journals such as Nature have had problems with statistical assessment of data over the years. In an editorial in April 2013, the editors state [2] "*Over the past year, Nature has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.*" Chemometrics suffers from the same problem [3].

The intellectual foundation of chemometrics was established in the 1970s, and most mainstream packages have retained this philosophy, despite areas such as applied statistics and machine learning moving in different directions. This is primarily because the methods are embedded in packages and there is considerable commercial advantage to selling these historic methods. Yet much of modern chemometrics has moved away from mainstream analytical chemistry to pattern recognition and hypothesis based studies, for example in heritage studies or metabolomics or medicine.

Partial Least Squares is one of the most overused and misunderstood methods. In classification, PLS-DA is widely used because of its availability in packages. PLS-DA with one component is identical to the well established method of Euclidean Distance to Centroids, and with all non-zero components to Linear Discriminant Analysis, dependent on data scaling. PLS-DA requires a large number of decisions, such as data scaling, decision thresholds, and when there are more than two groups, whether to use PLS1-DA or PLS2-DA. All the underlying decisions can make radical difference to the results, PLS really should be regarded as one step in many, rather like multiplication or division, but a bit more sophisticated, rather than a method in its own right, but very few users of this approach understand this. PLS is more relevant in calibration, although under certain circumstances this too is identical to other methods such as Principal Components Regression.

The difference between validation and optimisation is often badly misunderstood, probably because of the widespread historic use of cross-validation.

There is a bad tendency to compare methods in the literature, a method with "slightly higher" percent correctly classified or "slightly lower" error being considered better than another method, but the performance of methods depends crucially on data structure, approach to validation, relative class sizes (for classification), foreknowledge, pre-processing etc. And for real world situations, a low apparent error rate is not necessarily good, it can often be a result of over-fitting: the only sure way for comparison is using simulations and these in themselves often are not relevant to the case study in practice because we do not know everything in advance. Most statisticians now use Bayesian approaches, but these have been slow to take off in chemometrics, possibly because probabilities are hard to incorporate into algorithms such as PLS.

Few people check their data for normality, and yet many key statistics depend on this. Critical thresholds such as 95% or 99% cut-offs, to have much meaning, depend on data being normal "in the wings" something that is exceptionally difficult to test experimentally: most symmetric distributions are normal in the centre, but most decisions such as whether an observation is a member of a predefined group, depend on modelling the wings well (fig.1). Sample sizes are often grossly inadequate for this purpose.

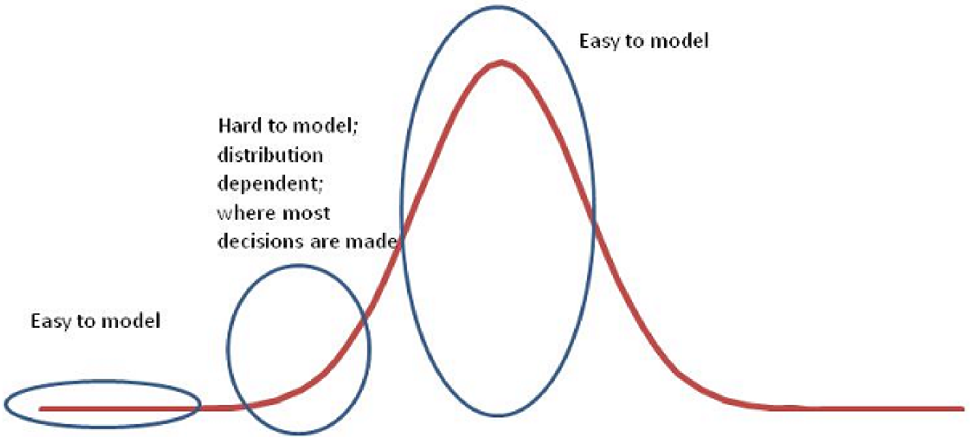

Fig.1 Wings of different distributions are hard to model but are usually key to decision making

Over-fitting is very common. PLS-DA scores plots on a training set, or even using a test set if variable selection is done in advance on the original data, will almost always show an apparent separation between classes even on completely random data. This is analogous to tossing a coin 10 times, occasionally there will be 8 or 9 Heads, so repeating this 1000 times and selecting the sets of tosses in advance, could give an impression that the coin is biased, when it is not. Many chemometrics packages are now sold to provide an overoptimistic view of the data which the market likes and therefore potentially misleads the user. PLS and its enhancements are particularly guilty of this. The use of null datasets or permutations of the classifier are rarely reported, and many suppliers of software do not like this as it may falsely suggest to the customer that their approach is "worse" than a competing approach.

Conclusion

Many other examples are widespread in the literature. Although the foundations in the 1970s were sound and advocated by careful scientists, the application areas have changed substantially, and with a much wider user base, few are wary of all the caveats. Finally software salesman and grant funded scientists alike want to present results in the most optimistic possible way even though the results may be statistically questionable.

References

1) D.L. Vaux, Research methods: Know when your numbers are significant, *Nature*, 492(7428), 2012, 180-181.
2) Editorial*,* Reducing our irreproducibility*, Nature*, 496(7446), 2013, 398.
3) R.G.Brereton, A short history of chemometrics: a persona view, *J Chemometrics*, 28(10), 2014, 749-760. DOI: 10.1002/cem.2633