This was originally published in Spectroscopy, 14(6), p. 12-14 (1999)

# Statistics in Spectroscopy Part 33 - Linearity in Calibration - Act II Scene V

by H. Mark and J. Workman

This column is still a continuation of our discussion started by the responses received to our column “Linearity in Calibration”(1). So far our discussion has extended over these previous columns:(2-5). At this point, however, we are finally getting toward the end of our obsession with considerations of linearity - at least until we receive another set of comments from our readers. Incidentally, we welcome such feedback, even those that disagree with us or with which we disagree, so please keep it coming. Indeed, it seems that we don’t get much feedback unless our readers disagree with us, and feel it strongly enough to have the need to say so. That’s great - there’s nothing like a little controversy to keep a column like this interesting: who said chemometrics and statistics and mathematics were dry subjects, anyway?!

In our original column on this topic(1) we had only done a principal component analysis to compare with the MLR results. One of the comments made, and it was made by all the responders, was to ask why we didn’t also do a PLS analysis of the synthetic linearity data. There were a number of reasons, and we offered to send the data to any or all of the responders who would care to do the PLS analysis and report the results. Of the original responders, Paul Chabot took us up on our offer. In addition, at the 1998 International Diffuse Reflectance Conference (The “Chambersburg” meeting), Susan Foulk also offered to do the PLS analysis of this data.

Gratifyingly, when Paul and Susan reported their PLS loadings they were identical, even though they used different software packages to do the PLS calculations (PLSIQ and Unscrambler). We’re certainly glad we don’t have to worry about sorting out differences in software packages (due to different convergence criteria, etc., that sometimes creep into results such as these) on top of the Chemometric issues we want to address.

Figure 1 presents the plot of the PLS loadings. Paul and Susan each computed both loadings. Note that the first loading is indistinuishable to the eye from the first PCA loading (see our original column on this topic(1)).

Paul and Susan each also computed the two calibration models and performance statistics for both models. Except that various programs didn’t compute the same sets of performance statistics (although in one case a different computation seemed to be given the same label as SEE), the ones that were reported by both programs had identical values.

As expected by all responders, and by your hosts as well, when two-factor models (either PCR or PLS) were computed, the fit of the model to the synthetic data was perfect. Table 1 presents as summary of the numerical results obtained, for one-factor calibration models.

Interestingly, when comparing the calibration results we find that the reported correlation coefficients agree among the different programs using the same algorithm, but the SEE values differ appreciably; it would seem that not all programs use the same definition of SEE. This leaves in question, for example, whether the value reported for SEE from PLS by Susan Foulk is really as large an improvement over the SEE for PCR reported by your columnists, or if it is due to a difference in the computation used. Since Paul Chabot reported SEE for both algorithms and his values are more nearly the same, even though his computation seems to differ from both the others, the tentative conclusion is that there is a difference in the computation. Indeed, we find that if we multiply our own value for SEE by the square root of 4/5 we obtain a value of 0.0514045, a value that compares to the SEE obtained by Susan Foulk in more nearly the same way that Paul Chabot's values compare to each other, indicating a possibility that there is a discrepancy in the determination of degrees of freedom that are used in the two algorithms.

Based on the values of the correlation coefficients, then, we can find the following comparisons between the two algorithms: as several of the responders indicated, the PLS model did provide improved results over the PCR model. On the other hand, the degree of improvement was not the major effect that at least some of the responders expected. As Richard Kramer expected:

> "PLS should have, in principle, rejected a portion of the non-linear variance resulting in a better, although not completely exact, fit to the data with just 1 factor. ... "

Some of this variance was indeed rejected by the PLS algorithm, but the amount, compared to the Principal Component algorithm seems to have been rather minuscule, rather than providing a nearly exact fit.

Non-linearity is a subject the specifics of which are not prolifically or extensively discussed as a specific topic in the multivariate calibration literature, to say the least. Textbooks routinely cover the issues of multiple linear regression and non-linearity, but do not cover the issue with "full-spectrum" methods such as PCR and PLS. Some discussion does exist relative to multiple linear regression, for example in "Chemometrics: a Textbook" by D.L Massart et al.,(6), see Section 2.1: Linear Regression (pp. 167-175) and Section 2.2: Non-linear Regression (pp. 175-181). The authors state:

> "In general, a much larger number of parameters {wavelengths, frequencies, or factors} needs to be calculated in overlapping peak systems {some spectra or chromatograms} than in the linear regression problems" (p. 176).

The authors describe the use of a Taylor expansion to negate the second and higher order terms under specific mathematical conditions in order to make "any function" (i.e., our regression model) first-order (or linear). They introduce the use of the Jacobian matrix for solving non-linear regression problems and describe the matrix mathematics in some detail (pp. 178-181).

There are also forms of non-linear PCR and PLS where the linear PCR or PLS factors are subjected to a non-linear transformation during singular value decomposition; the non-linear transformation function can be varied with the non-linearity expected within the data. These forms of PCR/PLS utilize a polynomial inner relation as spline fit functions or neural networks. References for these methods are found in (7). A mathematical description of the non-linear decomposition steps in PLS is found in (8).

These methods can be used to empirically fit data for building calibration models in non-linear systems. The interesting point is that there are cases, such as the one demonstrated in the Linearity in Calibration column where nonlinearity is the dominant phenomenon, where MLR will fit the data more closely with fewer terms than either PCR or PLS. One could imagine a real case where an analyte would have a minor absorption band such that the magnitude of the spectral band is within a linear region of the measuring instrument. One could also imagine the major absorption band of this analyte is somewhat non-linear at the higher concentration ranges. In this special case the MLR would provide a closer fit with fewer terms than either the PLS or PCR, unless the minor band was isolated prior to model development using the PCR or PLS. This points to a continuing need for spectral band selection algorithms that can automatically search for the optimum spectral information and linear fit prior to the calibration modeling step. But all things remaining constant, cases remain where MLR with automatic channel selection feature will provide a more optimum fit, in some cases, than either PCR or PLS. Surprising indeed, to some people!

In their day, Principal Components and Partial Least Squares were considered almost as “the magic answer to all calibration problems”, It took a long time for the realization to dawn that they contain no “magic” and are subject to most of the same problems as the algorithm previously available (at that time, what we now call MLR). Now we see a surge in other new algorithms: wavelets, neural networks, genetic algorithms, as well as the combining of techniques (e.g., selecting wavelengths before performing a PCA or PLS calculation). While some of the veterans of the “PC wars” (not “political correctness”, by the way) realize that they can be overfit just as MLR calibrations can, have become wary of the problem and are more cautious with new algorithms, there is some evidence that a large number, perhaps the majority, of users are not nearly so careful, and are still looking for their “magic answer”. There is a generic caution that need to be promoted, and all users made aware of when dealing with these more sophisticated. That is the simple fact that every new parameter that can be introduced into a calibration procedure is another way to overfit and hide the fact that it is happenning. Worse, the more sophisticated the algorithm the harder it is to see and recognize that that is going on.

With PCR and PLS we introduced the extra parameter of the number of factors: one extra parameter. With wavelets we introduce the order and the locality of each wavelet: two extra parameters. With neural nets, we have the number of nodes in each layer: n extra parameters, and then there is even a metaparameter: the number of layers. No wonder reports of overfitting abound (and don’t forget: those are only the ones that are recognized)!

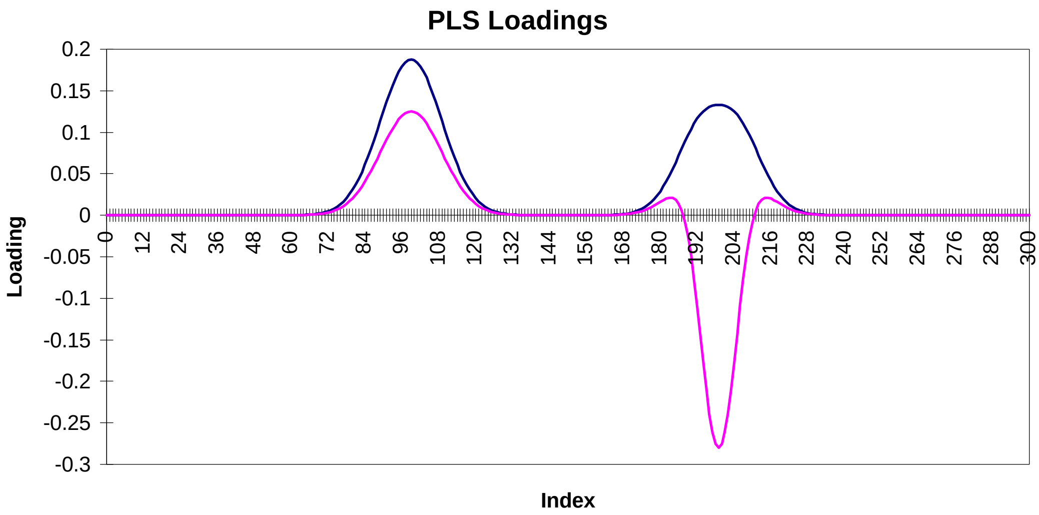
And nary a diagnostic in sight.

In a perfect world, a new algorithm would not be introduced until a corresponding set of diagnostic methods were developed to inform the user how the algorithm was behaving. As long as we’re dreaming let’s have those diagnostics be informative, in the sense that if the algorithm was misbehaving, it would point the user in the proper direction to fix it.

Table 1 - summary of results obtained from synthetic linearity data using one PCA or PLS factor. We present only those performance results listed by the data analyst as Correlation Coefficient and Standard Error of Estimate:

| Data analyst | Type of analysis | Corr. Coeff. | SEE |
|---|---|---|---|
| Column | PCR | 0.999622439 | 0.057472 |
| Chabot | PCR | 0.999622411 | 0.01434417 |
| Chabot | PLS | 0.999623691 | 0.01436852 |
| Foulk | PLS | 0.999624 | 0.051319 |

## Figure 1


PLS Loadings
Loading
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
0
12
24
36
48
60
72
84
96
108
120
132
144
156
168
180
192
204
216
228
240
252
264
276
288
300
Index

# REFERENCES

1. Mark, H., Workman, J.; Spectroscopy; 13 (6), p.19-21 (1998)

2. Mark, H., Workman, J.; Spectroscopy; (1998)

3. Mark, H., Workman, J.; Spectroscopy; (1998)

4. Mark, H., Workman, J.; Spectroscopy; (1998)

5. Mark, H., Workman, J.; Spectroscopy; (1998)

6. Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., Kaufman, L.; "Chemometrics: a Textbook"; Elsevier Science Publishers; Amsterdam (1988)

7. Wold, S., Kettanah-Wold, N., Skagerberg, B.; Chemom. Intell. Lab. Syst.; 7 p.53-65 (1989)

8. Wold, S.; Chemom. Intell. Lab. Syst.; 14 (1992)