This was originally published in Spectroscopy, 14(2), p. 16-27, 80 (1999)

Statistics in Spectroscopy Part 31 - Linearity in Calibration - Act II Scene III

by H. Mark and J. Workman

Some time ago we wrote a column titled “Linearity in Calibration”(1), in which we presented some unexpected results when comparing a calibration model using MLR with the model found using PCR. That column generated a rather active response, so we are discussing the subject and responding to the comments received in some detail, spread over several columns. The first two parts of our response have recently been published(2, 3); this column is the continuation of those.

We ended the last column with a summary of the comments received regarding the “Linearity in Calibration” column. We therefore pick up where we left off by starting this column with that same summary (naturally, anyone who wishes to read the full text of the comments will have to go back and reread the previous column(3)):

1) Richard Kramer, Patrick Wiegand and Fred Cahn felt that we should have tried two factors.
2) Richard Kramer and Patrick Wiegand thought we should have added simulated noise to the data.
3) All four responders indicated that we should have tried PLS.
4) Richard Kramer, Patrick Wiegand and Paul Chabot indicated that one PLS factor might do as well as one wavelength.
5) Richard Kramer and Patrick Wiegand thought that our conclusion was the MLR is better than PCA.

In addition, each of the responders had some of their own individual comments; we discuss all these below.

We now continue with our responses, and discussion of these comments: It may surprise some to hear this, especially in the light of some of the comments we make below, but we agree with the responders more than we disagree. We also believe, for example, in pre-screening the data, at least as strongly as Patrick Wiegand does, and we believe his comments regarding the way all (or at least, let's hope all) experienced chemometricians approach a problem. Indeed, fully half the book that one of us authored (4) was spent on just that point: how to "look at the data". However, our experience in the "real world" (as some like to call it) of instrument manufacturers has given us a somewhat different slant on the reality of what actually happens when users get hold of a new super-whiz-bang package of calculation.

In 16 years of experience in the NIR applications department at Technicon Instruments, there was about an hour and a half available to teach both theory and practice of calibration to each group of new users; the rest of the training time was spent teaching the students how to set the instrument up, prepare samples, take reproducible readings, and learn the rest of the mechanics needed to run the instrument, take readings and collect the data. How much attention do you think could be paid to the

finer points? This seems to be typical of what happens in the majority of cases involving novice users, and it is rare that there is anyone "back at the plant" who can pick up the ball and take them any further.

Even experienced practitioners can be misled, however. As was pointed out, real data contains various types and amounts of variations in both the X and Y variables. Furthermore, in the usual case, neither the constituent values nor the optical readings are spaced at nice, even, uniform intervals. Under such circumstances, it is extremely difficult to pick out the various effects that are operative at the different wavelengths, and even when the data analyst does examine the data, it may not always be clear which phenomena are affecting the spectra at each particular wavelength.

Now we will respond to the various comments, and make some more observations of our own. We will re-quote the pertinent parts of the communications from the responders, collecting together those on a similar topic and comment on them collectively. Note than some of these quotes were from later messages than those quoted in our previous column, because they were generated during subsequent discussions, and so may not have appeared previously.

We hope nobody takes our reply comments personally. Both some of the comments and some of our responses are energetic, because we seem to have touched on a subject that turned out to be somewhat controversial. So we do not take the responders comments personally, but we do enter with zest and gusto into what looks like something turning into a rather lively debate, and we sincerely hope that everybody can take our own comments in that same spirit.

The format of this columns is as follows: each numbered section starts with the comments from the various responders dealing with a given aspect of the subject, followed by our response to them collectively. So now let us consider the various points raised, starting with the use of noise-free data:

1. "You start with a 'perfectly noise-free spectrum' " (Patrick Wiegand)

"In regards to number 1, by using a perfectly noise-free spectrum, you have eliminated the main advantage of PLS/PCR. That is, the whole point of using these techniques is that they have better ability to reject noise than MLR. To come to an adequate conclusion as to the best performer, you should at least add an amount of random noise an order of magnitude greater than normal, since the amount of nonlinearity you use is an order of magnitude greater than normal." (Patrick Wiegand)

"The second problem is that that we never have the luxury of working with noise-free data. Thus, the column did not ask the right question(s). The proper question to ask is 'In what ways and under which circumstances do the signal averaging advantages of multiple-wavelength models outperform or underperform with respect to a single (or n wavelength, where n is a small integer) wavelength calibration when noise is present?' The answer will depend upon the levels of noise and non-linearity and the number of wavelengths in each model." (Richard Kramer)

“It isn't a case of 'extreme difficulty'. It is a situation where, in one case you use a factor which happens to be based upon an explicit model (i.e. linearity) which is correct

for the data while stacking the deck against the second case by denying any opportunity to be correct." (Richard Kramer)

Response: Of course we used noise-free data. Otherwise we could not be sure that the effects we see are due to the characteristics we impose on the data, rather than the random effects of the noise. When anyone does an actual, physical, experiment and takes real readings, the noise level, or the signal-to-noise ratio is a consideration of paramount importance, and any experimenter normally takes great pains to reduce the noise as much as possible, for just that reason. Why shouldn't we do the same in a computer experiment?

On the other hand, PCA and PLS are both known to perform better than MLR when the data is noisy because of the inherent averaging that they include. In this we agree fully; indeed, we also mentioned this characteristic in the original column. Richard Kramer hit the nail on the head with his question "In what ways ...?" The important question, then, that needs to be asked (and answered) is: at what point does one phenomenon or the other become dominant, so as to control or determine which algorithm will provide a better model? The next important question is: how can we tell which phenomenon is dominant in any particular case?

Rich Kramer also had the insight to go to the next step, and realized that the only way to determine whether the non-linearity is "small" or "large" is by having something to compare to, and the natural characteristic to compare it to is the noise. On this score we also agree with Richard and Patrick fully, and this is one place where much research is needed (there are others; and we will get to them in due course): How do you compare the systematic behavior of non-linearity with the random behavior of noise? The standard application of the science of Statistics provides us with tools to detect systematic effects, but how do we go to the next step and ascertain their relative effects on calibration models? These are among the fundamental behavioral properties of calibrations that are not being investigated, but need to be.

There are important theoretical reasons to reduce the spectral noise when doing calibrations. Nevertheless, if the main advantage of PLS is its behavior in the presence of noisy data (as Patrick Wiegand states), that is poor praise indeed. Noise levels of modern instruments are far below those of the past. In some cases, and NIR instruments come to mind here, the noise levels are so low that they are tantamount to having "zero noise" to start with. This improvement in instrumentation is a good thing, and we sincerely doubt that anybody would recommend using a noisy instrument for the sole purpose of justifying a more sophisticated algorithm.

In any case, even if all the above statements are 100% true, it does not affect our discussion because they are beside the point. The behavior of calibration algorithms in the face of noisy data is an important topic and perhaps should be studied in depth, but it was not at issue in the "Linearity in Calibration" column.

2. "You create an excessively high degree of non-linearity which would never be tolerated by an experienced spectroscopist." (Patrick Wiegand)

Response: In the absence of random variation, ANY amount of non-linearity would give the same results, and if we used less, any differences from the results we presented would be only of degree, not of kind. Any amount of non-linearity is infinitely greater than zero. As we explained in the original column, we deliberately chose an unrealistically large amount of non-linearity for pedagogical purposes; what would be the point of comparing different calibration lines that the naked eye saw as equally straight? The fact that it is "unrealistically" large is immaterial.

3. "You assume the spectroscopist will use the entire spectrum blindly when applying PLS or PCR, even though some parts of the spectrum clearly have no information and other parts are clearly nonlinear." (Patrick Wiegand)

Response: above, I described the situation as we see it, regarding the traps that both experienced and novice users of these very sophisticated algorithms can fall into. Keep in mind the pedagogy involved as well as the chemometrics: by suitable choice of values for the "constituent", the peaks at the nonlinear wavelengths could have been made to appear equally spaced, and the linear wavelengths appear stretched out at the higher values. The "clarity" of the non-linearity is due to the presentation, not to any fundamental property of the data, and this clarity does not normally exist in real data. How is someone to detect this, especially if not looking for it? Attempts to address this issue have been made in the past (see (5)) with results that in our opinion, are mixed, at best. And that simulated data was also noise-free.

With real data, a more scientifically valid approach would be to correct the non-linearity from physical theory. In the current case, for example, a scientifically valid approach would be to convert the data to transmission mode, subtract the stray light and reconvert to absorbance: the nonlinear wavelengths would have become linear again. There are of course, several things wrong with this procedure, all of them stemming from the fact that this data was created in a specific way for a specific purpose, not necessarily to be representative of real data:

A) You would have to know a priori that only certain wavelengths (and which ones) were subject to the "stray light" or whatever source of nonlinearity was present.

B) One of the problems of current chemometric practice is the "numbers game" aspect. No matter how soundly based in physical theory a procedure is, if the numbers it produces are not as good (whatever that might mean in a specific case) as a different, more empirical, procedure, the second procedure will be used, no matter how empirical its basis. The counter-argument to that, of course, is something on the order of "Well, we have to get as good results as we can for the user" and there is a certain amount of legitimacy to this statement. However, we know of no other field of scientific study where a situation of this sort is tolerated. Certainly, every field has areas of unknown effects where not all the fundamental physical theory is available, but in all fields other than chemometrics, there are workers investigating these dark areas, to try to fill in the missing knowledge. In chemometrics, on the other hand, for at least the twenty-two years we have been involved with the field, all we have seen the workers in the field doing are building bigger and higher and more fanciful mathematical superstructures on foundations that few, if any of them, seem to be aware of. We will have more to say about this below.

C) The simple fact that sometimes, the nature of the correct physical theory to use in unknown.

D) Finally, the real reason we presented these results the way we did was that the whole purpose of the exercise was to study the effect of this type of variation of the data, so that simply removing it would not only be trivial, it would also be a counterproductive procedure.

4. "If I understand the column correctly, a 1-factor model was used. Well, a single linear factor can never be sufficient to properly model a non-linear system. A minimum of 2 factors are required." (Richard Kramer)

"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. ... The PLS does tend to reject (exclude) those portions of the x-data which do not correlate linearly to the y-block." (Richard Kramer)

“You limit the number of factors for PLS/PCR to 1, even though the number of latent variables must be greater, due to the nonlinearity.” (Patrick Wiegand)

"In principle, in the absence of noise, the PLS factor should completely reject the non-linear data by rotating the first factor into orthogonality with the dimensions of the x-data space which are 'spawned' by the non-linearity. The PLS algorithm is supposed to find the (first) factor which maximizes the linear relationship between the x-block scores and the y-block scores. So clearly, in the absence of noise, a good implementation of PLS should completely reject all of the non-linearity and return a factor which is exactly linearly related to the y-block variances." (Richard Kramer)

"While I am no longer working in this field, and cannot easily do simulations, I think that a 2 factor PCR or PLS model would fully model the simulated spectra." (Fred Cahn)

"My "objection" is that you did not seem to look at the 2nd factor, which I think is needed to accurately model the spectra after the background is added." (Fred Cahn)

"I would expect PLS to outperform PCR, and the loading of the first principal component to be mostly located around the lower wavelength peak for PLS." (Paul Chabot)

Response: Yes, but.

The point being that, as our conclusions indicate, this is one case where the use of latent variables is not the best approach. The fact remains that with data such as this, one wavelength can model the constituent concentration exactly, with zero error - precisely because it *can* avoid the regions of non-linearity, which the PCA/PLS methods cannot do. It is not possible to model the "constituent" better than that, and even if PLS could model it just as well (a point we are not yet convinced of since it has not yet been tried - it should work for a polynomial non-linearity but this nonlinearity is

logarithmic) with one or even two factors, you still wind up with a more complicated model, something that there is no benefit to.

Richard Kramer suggested that we use two wavelengths (with the MLR approach) to see what happens. Well, here's what happens: if the second wavelength is also on the linear absorbance band, you get a "divide by zero" error upon performing the matrix inversion due to the perfect collinearity between the data at the two wavelengths. If the second wavelength is on the nonlinear band, the regression coefficient calculated for it is exactly zero (at least to 16 digits, where the computer truncation error becomes important), since it plays exactly no role in the modeling. In other words, not only is it unnecessary to add a second wavelength to the model, it is impossible to do so if you try; when the model is perfectly correct you can't force a second wavelength into that model even if you want to.

Richard Kramer, Patrick Wiegand and Paul Chabot suggested that a one-factor PLS model should reject the data from the nonlinear wavelength and therefore also provide a perfect fit to the "constituent". I offered to provide the data as an EXCEL spreadsheet to these responders; Paul accepted the offer, and I e-mailed the data to him. We will see the results at an appropriate stage.

5. "There are many well-established techniques for choosing which wavelength regions to use when modeling with PLS/PCR. First, I advise people to make sure that the pure component spectrum actually has a band in the location being modeled." (Patrick Wiegand)

Response: That indeed is a good procedure when you can do it (keeping in mind our earlier discussion regarding users reactions to the case of a conflict between theoretical correctness and the experimental “numbers game”), and we also make the same recommendation when appropriate. If anything, proper wavelength choice is even more important when using MLR than either PCA or PLS. But what do you do when the "constituent" is a physical property, with no distinct absorbance band? This consideration becomes particularly pernicious when that property is not itself being calibrated for, but is a variation superimposed on the data, and needs a factor (or wavelength) to compensate for, yet has no absorbance band of its own? The prototype example of this is the "repack" effect found when the measurements are made by diffuse reflectance: "Repack" does not have an absorbance band.

Other situations arise where that approach fails: when the chemistry is unknown or too complicated (octane rating in gasoline, for example). Here again, even though a fair amount is known about the chemistry behind octane rating, there is no absorbance band for "octane value".

Another case is where the chemistry is known, but the spectroscopy is unknown, because the pure material is not available. Protein, for example, cannot be extracted from wheat (or at least not and still remain protein), so the spectrum of "pure" protein as it exists in wheat is unknown. Even simpler molecules are subject to this effect: we can measure the spectrum of pure water easily enough, for example, but that is not the same spectrum as water has when it is present as an intimate mixture in a natural

product - the changes in the hydrogen bonding completely change the nature of the spectrum.

And these examples are ones we know about!

6. "Finally, the calibration statistics presented in Table 1 show a correlation coefficient of 0.9996 for PCR, even when an obviously nonlinear region is used! I am not sure if this is significantly different from the 1 shown for MLR using only the linear region. To me either model would be acceptable at the stage of method development where the article ended. Besides, it is unlikely that someone would be able to know a priori that the linear region was the better region to use for MLR." (Patrick Wiegand)

Response: As a purely practical matter, we agree with that interpretation. However, we hope that by now we have convinced you that we are trying to do more than that - we are trying to find out what really goes on inside the "black boxes" of chemometric calculations. The fact that the value of the PCR correlation coefficient differs significantly from unity becomes clear when you look at the other term of the ANOVA equation: in the MLR case the sum-squared error is zero, in the PCR case it is "infinitely" greater than that. Don't forget that "significance", at least in the statistical sense, is defined only when dealing with random variables. This also relates to the earlier comment regarding how to find ways to compare the relative effects of noise and non-linearity on calibration models.

7. "It would be very interesting also, since the performance of the models presented are so similar, to see how the performance would be affected by noise, drift, etc. which are always present in actuality. I would not be surprised if PLS/PCR outperformed MLR under those circumstances. (Patrick Wiegand)

Response: Yes, it certainly would be most interesting to investigate this question. This is closely related to the previous discussion concerning the relationship between noise and non-linearity, so I would modify the statement of the problem to "At what point does one or another effect dominate the behavior of the calibration?" i.e., where is the crossover point? Investigating questions of this sort is called “research”, and a more fundamental question arises: why isn’t anybody doing such investigations?

Other, related, questions are also important: Having determined this in isolation, how does the data analyst determine this in real data, where unknown amounts of several effects may be present? There a similarity here to Richard's earlier point regarding the relationship between the amount of noise and the amount of non-linearity. Here are more fertile areas for research into the behavior of calibration models.

8. "At any wavelength in your simulation, a second degree power series applies, which is linear in coefficients, and the coefficients of a 2 factor PCR or PLS model will be a linear function of the coefficients of the power series. (This assumes an adequate number of calibration spectra, that is, at least as many spectra as factors and a sufficient number of wavelength, which the full spectrum method assures.) The PCR or

PLS regression should find the linear combination of these PCR/PLS coefficients that is linear in concentration." (Fred Cahn)

Response: We have read the indicated section of that paper (6), and scanned the rest of it. We agree with much of what it says, both in the paper and in Fred Cahn's messages, but we're not sure we see the relevance to the column.

Certainly, non-linearities in real data can have several possible causes, both chemical (e.g., interactions that make the true concentrations of any given species different than expected or might be calculated solely from what was introduced into a sample, and interaction can change the underlying absorbance bands, to boot) and physical (such as the stray light, that we simulated). Approximating these non-linearities with a Taylor expansion is a risky procedure unless you know a priori what the error bound of the approximation is, but in any case it remains an approximation, not an exact solution. In the case of our simulated data, the non-linearity was logarithmic, thus even a second-order Taylor expansion would be of limited accuracy.

Alternative methods, such as correcting the nonlinearity though the application of an appropriate physical theory as we described above, may do as well or even better than a Taylor series approximation, but a rigorous theory is not always available. Even in cases where a theory exists, often the physical conditions for which the theory is valid cannot be achieved; we demonstrated this in the discussion in a previous column(2) of the fundamental impossibility of truly achieving "Beer's Law linearity".

Thus we are left with a situation where even in the best cases we can achieve, there can be residual non-linearities in the data. The purpose of our column was to investigate the behavior of different modeling methods in the face of non-linearity.

9. "Thus, my interest in 2 or more factor chemometric models of your simulation is in line with this view of chemometrics. I agree with the need for better physical understanding of instrument responses as well as of the spectra themselves. I would not choose PCR/PLS or MLR to construct such physical models, however." (Fred Cahn)

Response: We were not trying to use the chemometric techniques to create a physical model in the column. We also agree that physical models should be created in the traditional manner: based on the study of the physical considerations of a situation. Ideally you would start from a fundamental physical law and derive, through logic and mathematics, the behavior of a particular system: this is how all other fields of science work. A chemometric technique then would be used only to ascertain the value (from a series of physical measurements) of an unknown parameter that the mathematical derivation created.

What we were trying to do in the column was to ascertain the behavior of a mathematical (not physical!) system in the face of a certain type of (simulated) physical behavior. There is nothing wrong with trying to come up with empirical methods for improving the practical performance of chemometric calibration, but one of the philosophical problems with the current state of chemometrics is that nobody but

nobody is trying to do anything else: i.e., to determine the fundamental behavior of these mathematical systems.

10. “The synthetic data did NOT demonstrate the advantage of a single linear wavelength over a multiple wavelength {sic} model ...” (Richard Kramer)

" ... in one case you use a factor which happens to be based upon an explicit model (i.e. linearity) which is correct for the data while stacking the deck against the second case by denying any opportunity to be correct." (Richard Kramer)

“In your article, you appear to be creating an artificial set of circumstances: ...” (Patrick Wiegand)

"Thus your conclusion -- that MLR is more capable of producing accurate models than PLS/PCR -- is based on a contrived set of circumstances that would not occur in reality, especially when the chemometrician/spectroscopist is experienced." (Patrick Wiegand)

Response: Artificial? Contrived? Only insofar as any experimental study is based on a "contrived" set of circumstances - contrived to enable the experimenter to separate the phenomenon of interest and study its effects, with "everything else the same".

But that is a minor matter. Richard and Patrick (and how many others, who didn’t respond?) believe that we concluded that “MLR is better than PCA/PLS”. The really critical point here is that that is NOT our conclusion, and anyone who thinks it is has misunderstood us. We put the fault for this on ourselves, since the one thing that is clear is that we did not explain ourselves sufficiently.

Therefore let us clarify the point here and now: we are not fighting a "holy war" against PCA/PLS etc. The purpose of the exercise was NOT to "prove that MLR with wavelength selection is better", but to investigate and explain conditions that cause that to be so, when it happens (which it does, sometimes). As we discussed in the original column, more and more discussions about calibration processes, both oral and in the literature, describe situations where wavelength selection improved the results (in PCR and PLS as well as MLR), but there has previously been no explanation for this phenomenon. Therefore we decided to investigate non-linearity since we suspected that to be a major consideration, and so it turned out to be.

We continue our discussion in the following columns.

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.; "Principles and Practice of Spectroscopic Calibration"; John Wiley & Sons; New York (1991)

5. Mark, H.; Applied Spectroscopy; 42 (5), p.832-844 (1988)

6. Cahn, F., Compton, S.; Applied Spectroscopy; 42 p.865-872 (1988)