This was originally published in Spectroscopy, 14(1), p. 16-17 (1999)

Statistics in Spectroscopy Part 30 - Linearity in Calibration - Act II Scene II

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 an active response, so we are discussing the subject in some detail, spread over several columns. The first part of our discussions have recently been published(2); this column is the continuation of that one.

In this column we now present the responses we received to original one column(1) in order of receipt, following which we will comment about them in subsequent columns. Here, in order of receipt, are the comments:

> The first comments we received were from Richard Kramer:
>
> “Howard,
> I'm afraid that this month's Spectroscopy Column is badly off the mark (pun intended (with apologies)). The errors are two-fold with the most serious error so significant that the other error is moot.
> 1. 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. The synthetic data did NOT demonstrate the advantage of a single linear wavelength over a multiple wavelength model, it merely illustrated the fact that a single linear factor is not sufficient to model non-linear data. We could stop here, but, for the sake of completeness....
> 2. 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.
>
> Regards, Richard”

We went back and forth a couple of times, but rather than list each of our conversations individually, we will reserve comments until we have looked at all the comments, and then we will summarize our responses to all four respondents together, since several of these response comments say the same things, to some extent.

Second, we received comments from Patrick Wiegand:

> "Gents,
>
> I have always looked forward to reading your articles on Chemometrics in Spectroscopy. They are truly a valuable resource -- I usually cut them out and save them for future reference. However, I think your article "Linearity in Calibration" in the June 1998 issue of Spectroscopy leads the reader to an erroneous conclusion. This conclusion results largely because of the assumptions you make about the application of PLS and PCR.
>
> I know of no experienced practitioner of chemometrics who would blindly use the "full spectrum" when applying PLS or PCR. In the book by "Chemometrics" by Beebe, Pell and Seasholtz, the first step they suggest is to "examine the data." Likewise, Kramer in his new book has two essential conditions: The data must have information content and the information in the data must have some relationship with the property or properties which we are trying to predict. Likewise, in the course I teach at Union Carbide, I begin by saying that "no modeling technique, no matter how complex, can produce good predictions from bad data."
>
> In your article, you appear to be creating an artificial set of circumstances:
>
> 1. You start with a "perfectly noise-free spectrum"
>
> 2. You create an excessively high degree of non-linearity which would never be tolerated by an experienced spectroscopist.
>
> 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.
>
> 4. 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.
>
> 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.

Number 2 -- I understand that you wanted to use a high degree of nonlinearity so that the absorbance vs. concentration plot will be nonlinear to the naked eye, but you can't really expect to use this degree of nonlinearity to make a judgmental comparison between two techniques if it is not realistic that it will ever occur in real life.

Number 3 -- 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. If this is not possible, at least only include regions that look like valid bands -- no sense in trying to include low s/n baseline regions. Plots of a linear correlation coefficient vs. wavelength for the property of interest are also useful in choosing the right regions to include in the model.

Finally, if the initial model is built using the full-spectrum, an examination of factor plots would reveal areas in which there is no activity.

Number 4 -- In cases where there is no choice but to deal with nonlinearity in the spectra, then it will be necessary to use more factors than the number of chemical species in the system. Once again, an experienced practitioner will use other ways of choosing the right number of factors, like a PRESS plot, etc.

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. 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.

All of the above would seem to indicate that I am totally against using MLR. This is not the case. In my practice, I always try the simplest approach first. This means first trying MLR. If that does not work, then I use PLS. If that does not work -- well, some people may use neural networks, but I have not yet found a need to do so. I think you are right in saying that there has been a lot of hype over PLS (although not as much as there has been over neural nets!) In many cases MLR works great, and I will continue to use it. To paraphrase Einstein, 'Always use the simplest approach that works -- but no simpler.'"

The third comments we received were from Fred Cahn:

“I read your article in Spectroscopy (13(6), June 1998) with interest. However, I don't agree with the conclusions and the way your simulation was carried out and/or presented.

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. 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.

See my publication:
Cahn, F. and S. Compton, "Multivariate Calibration of Infrared Spectra for Quantitative Analysis Using Designed Experiments", Applied Spectroscopy, 42:865-872 (July, 1988)."

Fred supplied a copy of the cited paper, and we read it. Again, the comments about it will be included among the general comments.

And finally, the fourth comments we received were from Paul Chabot:

"Hello Howard,

I recently read your column in the Spectroscopy issue of June 1998, which was dealing with "Linearity in Calibration". First, I have to tell you that I really like your monthly column. You do a good job at explaining the basics and more of many topics related to chemometrics, and "demistify" the subjects.

As an avid user of PLS, I was concerned when you were comparing MLR to PLS and PCR on your synthetic data set. Even though I agree with you that in some cases, MLR is a much better approach than PLS or PCR, sometimes the use of a full spectrum technique is essential. In this particular case, I do not doubt your results showing that MLR outperforms the full spectrum techniques because the data set was designed to do so. But out of the full spectrum techniques, 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. Did you notice any difference between PCR and PLS on this data set? I would appreciate it if you could let me know if you tried both approaches and the results you obtained so I don't have to regenerate the data.

Thank you very much, and keep up the good work,

Paul Chabot"

To summarize the comments (including ones presented during subsequent discussions, and therefore not included above):

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.

As stated in the introduction to this column, we present our responses in columns to follow.

REFERENCES

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

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