This column was originally published in Spectroscopy, 13(6), p. 19-21 (1998)

CHEMOMETRICS IN SPECTROSCOPY Part 27: Linearity in Calibration

by Howard Mark and Jerome Workman

Those who know us know that we've always been proponents of the approach to calibration that uses a small number of selected wavelengths. The reasons for this are partly historical, since we became involved in Chemometrics through our involvement in Near-Infrared spectroscopy, back when wavelength-based calibration techniques were essentially the only ones available, and these methods did yeoman's service for many years. When full-spectrum methods came on the scene (PCR, PLS) and became popular, we adopted them as another set of tools in our chemometric armamentarium, but always kept in mind our roots, and used wavelength-based techniques when necessary and appropriate, and we always knew that they could sometimes perform better than the full spectrum techniques under the proper conditions, despite all the hype of the proponents of the full-spectrum methods. Lately, various other workers have also noticed that eliminating "extra" wavelengths could improve the results, but nobody (including ourselves) could predict when this would happen, or explain or define the conditions that make it possible. The advantages of the full-spectrum methods are obvious, and are promoted by the proponents of full-spectrum methods at every opportunity: the ability to reduce noise by averaging data over both wavelengths and spectra, noise rejection by rejecting the higher factors, into which the noise is preferentially placed, the advantages inherent in the use of orthogonal variables, and the avoidance of the time-consuming step of performing the wavelength selection process.

The main problem was to define the conditions where wavelength selection was superior; we could never quite put our finger on what characteristics of spectra would allow the wavelength-based techniques to perform better than full-spectrum methods.

Until recently.

What sparked our realization of (at least one of) the key characteristics was an on-line discussion of the NIR discussion group[i] dealing with a similar question, whereupon the ideas floating around in our heads congealed. At the time, the concept was proposed simply as a thought experiment, but afterward, the realization dawned that it was a relatively simple matter to convert the thought experiment into a computer simulation of the situation, and check it out in reality (or at least as near to reality as a simulation permits). The advantage of this approach is that simulation allows the experimenter to separate the effect under study from all other effects and investigate its behavior in isolation, something which cannot be done in the real world, especially when the subject is something as complicated as the calibration process based on real spectroscopic data.

The basic situation is illustrated in figure 1. What we have here is a simulation of an ideal case: a transmission measurement using a perfectly noise-free spectrometer through a clear, non-absorbing solvent, with a single, completely soluble analyte

dissolved in it. The X-axis represents the wavelength index, the y axis represents the measured absorbance. In our simulation there are six evenly spaced concentrations of analyte, with simulated “concentrations” ranging from 1 - 6 units, and a maximum simulated absorbance for the highest concentration sample of 1.5 absorbance units. Theoretically, this situation should be describable, and modeled by a single wavelength, or a single factor. Therefore in our simulation we use only one wavelength (or factor) to study.

For the purpose of our simulation, the solute is assumed to have two equal bands, both of which perfectly follow Beer’s law. What we want to study is the effect of non-linearities on the calibration. Any non-linearity would do, but in the interest of retaining some resemblance to reality, we created the nonlinearity by simulating the effect of stray light in the instrument, such that the spectra are measured with an instrument that exhibits 5% stray light at the higher wavelengths. Now, 5% might be considered an excessive amount of stray light, and certainly, most actual instruments can easily exhibit more than an order of magnitude better performance. However, this whole exercise is being done for pedagogical purposes, and for that reason, it is preferable for the effects to be large enough to be visible by eye; 5% is about right for that purpose.

Thus, the band at the lower wavelengths exhibits perfect linearity, but the one at the higher wavelengths does not. Therefore, even though the underlying spectra follow Beer’s law, the measured spectra not only show non-linearity, they do so differently at different wavelengths. This is clearly shown in Figure 2, where absorbance versus concentration is plotted for the two peaks.

Now, what is interesting about this situation is that ordinary regression theory, and the theory of PCA and PLS, specify that the model generated must be linear in the coefficients. Nothing is specified about the nature of the data (except that it be noise-free, as our simulated data is); the data may be non-linear to any degree. Ordinarily this is not a problem because any data transform may be used to linearize the data, if that is desirable.

In this case however, one band is linearly related to the concentrations and one is not; a transformation, blindly applied, that linearized the absorbance of the higher-wavelength band would cause the other band to become non-linear. So now, what is the effect of this all on the calibration results that would be obtained?

Clearly, in a wavelength-based approach, a single wavelength (which would be theoretically correct), at the peak of the lower-wavelength band would give a perfect fit to the absorbance data. On the other hand, a single wavelength at the higher-wavelength band would give errors due to the non-linearity of the absorbance. The key question then becomes: how would a full-wavelength (factor-based) approach behave in this situation?

In the discussion group, it was conjectured that a single factor would split the difference; the factor would take on some character of both absorbance bands, and would adjust itself to give less error than the non-linear band alone, but still not be as good as using the linear band.

Figure 3 shows the factor obtained from the Principal Component analysis of this data. It seems to be essentially Gaussian in the region of the lower-wavelength band, and somewhat flattened in the region of the higher-wavelength band, conforming to the nature of the underlying absorbances in the two spectral regions.

Because of the way the data was created, we can rely on the calibration statistics as an indicator of performance. There is no need to use a validation set of data here. Validation sets are required mainly to assess the effects of noise and intercorrelation. Our simulated data contains no noise. Furthermore, since we are using only one wavelength or one factor, intercorrelation effects are not operative, and can be ignored. Therefore the final test lies in the values obtained from the sets of calibration results, which are presented in Table 1.

Those results seem to bear out our conjecture. The different calibration statistics all show the same effects: the full-wavelength approach does seem to sort of “split the difference” and accommodate some, but not all, of the non-linearities, the algorithm uses the data from the linear region to improve the model over what could be achieved from the non-linear region alone.

On the other hand, it could not do so completely, it could not ignore the effect of the non-linearity entirely to give the best model that this data was capable of achieving. Only the single-wavelength model using only the linear region of the spectrum was capable of that.

So we seem to have identified a key characteristic of chemometric modeling, that influences the capabilities of the models that can be achieved: not non-linearity per se, because simple nonlinearity could be accommodated by a suitable transformation of the data, but differential, non-linearity, which cannot be fixed that way. In those cases where this type of differential, or non-uniform, nonlinearity is an important characteristic of the data, then selecting those wavelengths and only those wavelengths where the data are most nearly linear will provide better models than the full-spectrum methods, which are forced to include the non-linear regions as well, are capable of.

Now, the following discussion does not really constitute a proof of this condition (in the mathematical sense), but this line of reasoning is fairly convincing that this must be so. If, in fact, a full-spectrum method is splitting the difference between spectral regions with different types and degrees of non-linearity, then those regions, at different wavelengths, themselves must have different amounts of nonlinearity, so that some regions must be less nonlinear than others. Furthermore, since the full-spectrum method (e.g., PCR) has a non-linearity that is, in some sense, between that of the lowest and highest, then the wavelengths of least non-linearity must be more linear than the full-spectrum method and therefore give a more accurate model than the full-spectrum algorithm. All that is needed in such a case, then, is to find, and use those wavelengths.

Thus, when this condition of differential non-linearity exists in the data, modeling techniques based on searching through and selecting the “best” wavelengths (essentially we’re saying MLR) are capable of creating more accurate models than full-wavelength methods, since almost by definition this approach will find the

wavelength(s) where the effects of non-linearity are minimal, which the full-spectrum methods (PCA, PLS) cannot do.

Table 1 - Calibration statistics obtained from the three calibration models discussed in the text:

	Linear Wavelength	Non-linear Wavelength	Principal Component
SEE	0	0.237	0.0575
Corr. Coeff.	1	0.9935	0.9996
F	<infinite>	305	5294

FIGURE 1 - Six samples worth of spectra with two bands, without and with stray light

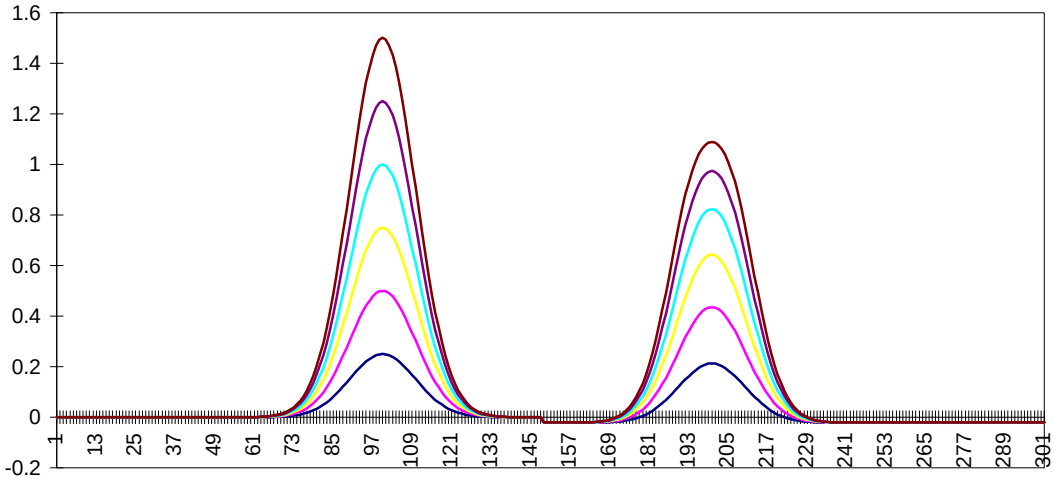


FIGURE 2 - Absorbance versus concentration, without and with stray light

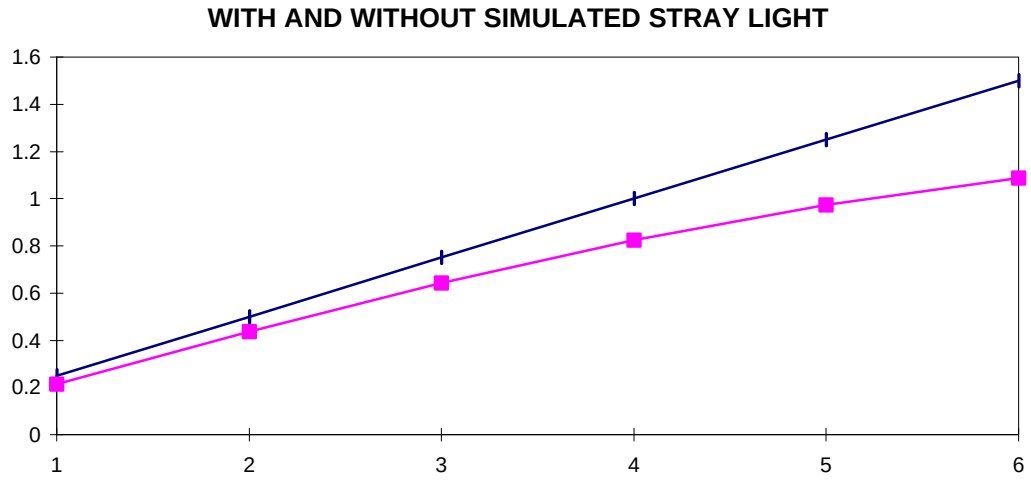
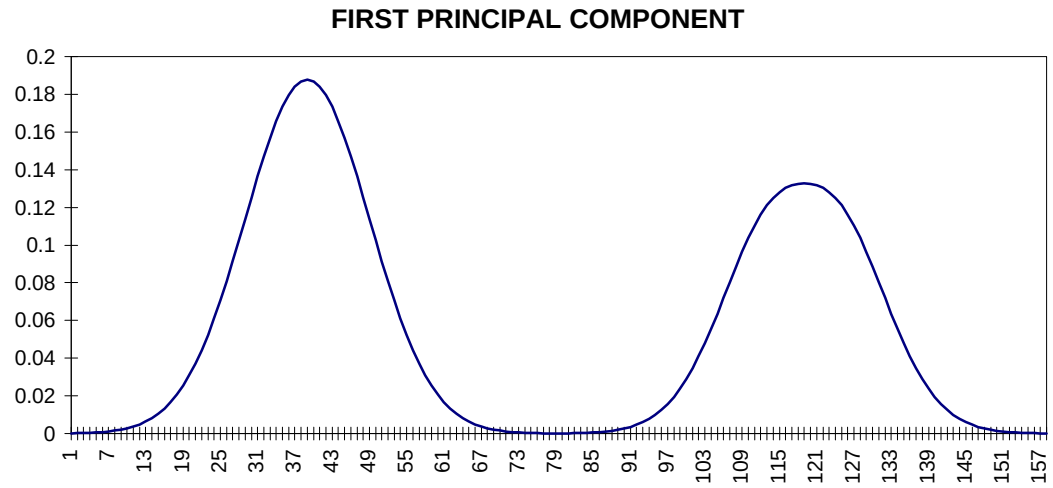


FIGURE 3 - First Principal Component from concentration spectra



[i] The moderator of this discussion group is Bruce Campbell. He can be reached for information, or to join the discussion group by sending a message to: CAMPCLAN@prodigy.net New members are welcome.