doi: 10.1255/nirn.1492

Eastern Analytical Symposium Award for outstanding achievements in near infrared spectroscopy: my contributions to near infrared spectroscopy

Mark Westerhaus
FOSS

At the recent EAS Symposium held in Somerset, NJ, USA, Mark Westerhaus was presented with the EAS Award for Outstanding Achievements in Near Infrared Spectroscopy. Mark kindly agreed to let us publish his talk so we are delighted to reproduce it below. I think you will find it to be a very clear explanation of the reasons for various elements of old and new software treatments that many of us have used in the past. Congratulations on this well-deserved award Mark!! Ed.

How I got started in NIR

I earned a BA and MA in psychology at Case Western Reserve University in 1974 and came to Penn State to study psychophysiology. While there, I started taking statistics courses and considered getting a dual degree. One term, the Psychology Department told me they were short on assistantships and requested that I ask the Statistics Department if they had an assistantship for me. I asked and they said a professor in the Agronomy Department was looking for someone to help with a project in near infrared (NIR) spectroscopy. I went for an interview and discussed the project with Dr John Shenk. I was hired by John on the spot and worked with him first at Penn State and later at Infrasoft International for the next 27 years until he retired.

I remember making two classic chemometric mistakes in my early career. The first involved an MLR program in FORTRAN we had received from Karl Norris. I noticed that it only allowed for 12 terms. I thought if 12 were good, 25 would be better and expanded the capability of the program. The fit results were great, but predictions on test samples were terrible. That was my first lesson in over-fitting!

Later, we wanted to create a model to predict the percentage of grass in a grass–legume mixture. We prepared one pile of grass and another pile of legume and carefully made mixtures for the calibration. Again we got excellent fit results, but terrible predictions when we tested the calibration with other grass and legume samples. Lesson 2: mixtures do not provide new information!

My contributions

I have picked six contributions that I believe can or will help the NIR user. Instrument standardisation and repeatability files enable a calibration developed at one site to predict more accurately at another site. Modified partial least squares (PLS) and LOCAL can produce more accurate calibrations than PLS. Global and Neighbourhood H statistics can be used to build datasets, detect outliers and identify groups of samples not covered by a calibration. More recently, Good Product Definition with Maximum Peak T was developed to detect contaminants in a product.

From left to right: Pierre Dardenne, Mark Wetserhaus, Anne-Françoise Aubry, Lars Nørgard and Søren Engelson.

Instrument standardisation

Calibration development at ISI followed the "one instrument, one lab" rule. While this rule made the calibration process simpler, it left us with the problem of transferring the calibration to another instrument as a separate issue. ISI developed two techniques to transfer a calibration—instrument standardisation and repeatability files.

The NIR instruments manufactured by NIRSystems did not output data at a constant wavelength spacing. Instead, the wavelength associated with the index of a transmitted data value was $k \times \sin(\text{position\_index} \times \pi / 7200 + \phi)$ for calculated values of k and ϕ. k and ϕ were calculated based on observed peak locations for internally-mounted materials with sharp absorption peaks. One problem was defining the nominal locations of the absorption peaks. ISI used the first set of nominals provided by NIRSystems and started building calibration datasets. NIRSystems tweaked the nominals several times in subsequent years to better match the average location observed on many instruments. ISI, however, continued to build calibration datasets based on the first set of nominals! That led to an

Table 1. Effect of Instrument Standardisation and Repeatability File for Calibration Transfer of Fat in Feed Samples.

	None	STD	Rep file	STD and rep file
SEP	1.796	0.338	0.352	0.225
Bias	1.774	0.024	0.265	0.057
SEP(C)	0.298	0.355	0.245	0.229

instrument setup decision on whether to have ISI compatibility or NIRSystems compatibility.

When ISI looked at many examples of the same model NIR instrument, we saw that there was often a spectral offset and/or scale difference between instruments, which we attributed to differences in the reference ceramic standards or light path geometries. We also noticed that small wavelength shifts were sometimes evident. When we compared a NIRSystems 5000 to a Technicon 500 instrument, we saw wavelength shifts that needed more than a linear correction. The ISI solution was multi-sample standardisation, in which 30 sealed samples are scanned on two instruments. Why 30? After a wavelength correction, a simple linear regression was used to get the scale difference at each wavelength. Also, 30 samples fit nicely in the metal box used to ship the sealed samples! A wide variety of products were used to ensure sufficient spectral variation for a slope to be calculated from the 30 pairs of values at each wavelength.

As instrument manufacturing improved, it became clear that the main difference between instruments was the offset. The ISI solution was single-sample standardisation, in which one sealed sample is scanned on two instruments and the difference between the scans used to correct the offset problem. The only question is which sealed sample to use. We found that, although the largest difference between instruments was an offset, other differences still existed. Applying an offset correction to a transfer problem that was not 100% offset worked perfectly for the transfer sample, but less well for samples that differed from the transfer sample. The solution was to make a single sample standardisation for each product, using a sample with a spectrum which was close to the average spectrum for that product. Instrument standardisation is becoming less important within a product line as manufacturers succeed in minimising instrument differences during production.

When comparing the two instruments, I scan a stable, sealed sample on both instruments and look at the difference spectrum. If sealed samples are not available, I use an average spectrum from a set of stable samples scanned on both instruments. If the difference spectrum looks like a small version of the spectrum, there is a scale difference between instruments. If the difference spectrum looks like the first derivative of the spectrum, there is a wavelength shift. If the difference spectrum looks like the second derivative of the spectrum, there is a bandwidth difference. Any remaining differences are assumed to be an offset difference.

Repeatability files

Even after applying instrument standardisation, there were often small differences remaining between instruments. These differences were large enough to cause a prediction bias when a calibration model developed on one instrument (the master) was used on another instrument. I wanted to modify the least squares solution to the regression problem to include minimising the prediction bias for the same sample scanned on the two instruments. In normal regression, the predicted value for the average spectrum is the average reference value. With a repeatability file, we want the average spectrum plus and minus the differences in the file to still predict close to the average reference value.

When using a repeatability file, the calibration error will be slightly larger, since the solution is no longer the least squares solution. However, the increase in calibration error is usually small compared to the improvement in repeatability. Repeatability files are an excellent way to prepare a calibration for variation not included in the calibration samples. Although developed for instrument variation, repeatability files are also an excellent way to remove temperature sensitivity from a calibration by adding groups of sample scans taken at different temperatures. Note that repeatability files are only needed for variation not already in the calibration samples.

Repeatability files have the advantage of not needing reference values. Also, there is no need to worry about having enough special samples to influence a large calibration set. As an example, a model for fat in feed created from Foss 6500 spectra was deployed on a Foss DS2500 (Table 1). The prediction statistics were computed with no transfer aid, single sample instrument standardisation only, recreating the calibration with a repeatability file and using both standardisation and a repeatability file. The best results were obtained by using both instrument standardisation and a repeatability file.

Modified PLS (MPLS)

I was trying to find a way to use squared correlation as weights to construct factors for multiple linear regression (MLR) when I heard about PLS. Once I learned that PLS was similar to what I was trying to do, I combined the two methods. Instead of using the $x'y$ covariance as a wavelength weight, I used the $x'y/x'x$ regression slope. Once I calculated the wavelength weights, I projected them onto the original spectra space. Otherwise, the standard PLS algorithm is used. PCA is designed to explain the most spectral variation possible with each factor. Each PLS factor is a compromise, trying to explain spectral variation while trying to correlate highly with the *Y* variable. Each MPLS factor is also a compromise, but places less emphasis on explaining the spectral variation.

Table 2. Prediction errors for PLS and MPLS for flour calibrations.

Constituent	Prediction error	
	PLS	MPLS
Protein	0.408	0.284
Ash	0.038	0.037
DM	0.196	0.176

If most of the spectral variation is due to the calibration constituent, PLS should do a good job. MPLS is useful when you need to ignore the major sources of spectral variation and concentrate on regions with smaller spectral variation but strong correlations. The one disadvantage I have found working with MPLS is the interpretability of the regression coefficients. They are not as smooth as PLS coefficients and are harder to relate to known constituent peaks. However, the coefficients for most PLS models with derivative pretreatment and many factors are also difficult to interpret.

As an example, a flour sample scanned on a Foss 5000 set was split on a time of analysis gap, calibrating on 1677 older samples and predicting 294 newer ones (Table 2). The math treatment was scale and offset correction followed by a first derivative (1,5,1,1). The wavelengths 1108–2492,2 were used. MPLS provided smaller prediction errors for all constituents.

LOCAL calibration

PLS can provide very accurate calibrations when constituents vary within a single product. One of the first problems we encountered where this was not the case was pasture analysis. Pasture was usually a mixture of grass and legume, with the mixture percentage varying widely. We saw accurate results when the mixture percentages were limited to a small range but the accuracy suffered when we tried to calibrate over a wider range. I developed LOCAL as a way to select an appropriate subset of the calibration samples for each sample being analysed.

LOCAL uses correlation to find samples similar to the one being analysed. It can use a derivative pre-treatment of the spectral data, but does not need scatter correction, since correlation ignores scale and offset differences. The number of selected samples to use is important. If too few are used, there may not be sufficient variation to create a good model. If too many samples are used, the problem may contain too much variation for a linear model like PLS to handle. There should be a sufficient number of samples in the LOCAL library so that the similar samples found for each sample being analysed are helpful in the calibration. LOCAL includes a batch option, in which several numbers of selected samples can be evaluated and compared automatically.

Once the similar samples have been identified, LOCAL performs a standard PLS calibration, storing information about each model from one factor to a pre-selected maximum number of factors. Multiple outlier elimination passes can be used to help ignore mistakes, either in the sample selection or the reference data. Instead of picking a model with a certain number of factors, LOCAL uses a weighted average of predictions from models that vary in the number of factors. During the training phase, all possible number of factors ranges are evaluated and the best ones are displayed. The best minimum and maximum number of factors are then used during routine analysis.

LOCAL should not be used to combine two dissimilar products into one LOCAL library. In the best case, the accuracy equals that of separate LOCAL libraries. In the worst case, an incorrect product is selected and the accuracy suffers. LOCAL does not make model validation easier. Each group of new samples should be monitored as if it had a unique calibration. A LOCAL calibration is often less repeatable that a static calibration. If a sample is re-scanned and re-predicted, the new set of selected similar samples may differ slightly from the previous set. The difference of even one calibration sample can cause a small but noticeable difference in the predicted value. Any loss in repeatability, however, is small compared to the gain in prediction accuracy.

As an example, 2109 poultry meal samples were scanned in a Foss InfraXact. Every fourth sample was reserved as a test

Table 3. Prediction errors for poultry meal using MPLS and LOCAL with test set constructedusing every fourth sample.

	Prediction error			
	Protein	Fat	Ash	Dry matter
MPLS	1.515	0.757	1.570	0.442
LOCAL	1.187	0.630	0.898	0.327
LOCAL GH	0.969	1.012	2.423	0.958
LOCAL NH	0.284	0.308	1.225	0.255

Table 4. Prediction errors for poultry meal using MPLS and LOCAL using the most recently scanned samples as the test set.

	Prediction error			
	Protein	Fat	Ash	Dry matter
MPLS	2.165	1.054	3.850	0.909
LOCAL	2.642	1.556	2.838	0.879
LOCAL GH	2.968	3.308	4.292	2.966
LOCAL NH	1.480	1.607	2.049	1.375

sample and the remaining were used as calibration samples. The spectra were pretreated with scale and offset correction and a first derivative (1,6,1,1). The test set was predicted using an MPLS calibration made on all calibration samples and LOCAL using 150 samples (Table 3). Note that the average Global H and average Neighbourhood H values are high for ash but as expected for the other constituents.

Another chemometrics lesson I have learned the hard way is that splitting a large group of samples randomly into calibration and test sets does not provide a realistic validation. It would be nice if it did, but all too often, time and location of samples being analysed differ from those in the calibration set. Whenever asked, I tell people that the best way to estimate future performance is to spit samples into an older and new group and use the older group to predict the newer group. If the samples came from several regions, it is also helpful to separate one or more regions into the test set.

Following that advice, I split the same samples set into 1272 older samples for calibration and 827 newer samples for testing. The particular numbers arose by splitting on a large time gap in the scan times. Other than the new split, all calibration parameters were the same as the previous test (Table 4).

The errors are roughly twice as large with this new split as they were with the every fourth sample split and the average LOCAL Global H and Neighbourhood H values are large enough to indicate a problem. Something changed over that time gap used for the split. Neither model was prepared for the change, but the H values gave an indication that there was a problem.

Global and neighbourhood H statistics

The "H" in Global H stands for Hat, and was inspired by "The Hat Matrix in Regression and ANOVA" by Hoaglin and Welsch, published in *The American Statistician*, 1978. Hat refers to the "^" symbol used to denote a fit value. In MLR regression, $Y\text{-hat} = X(X'X)^{-1}X'Y = \mathbf{H}Y$. The i^{th} diagonal element of the **H** matrix is the leverage for the i^{th} sample. The square root of the leverage value is also known as Mahalanobis distance. Global H is a leverage value scaled so that the average value in the calibration set is 1.0. It measures the squared distance of a spectrum to the average spectrum. When calibrating, the spectra with the largest GH values should be checked to see if they are mistakes to be removed from the calibration.

Global H can also be computed for new samples not included in the calibration; this is done by measuring the squared distance from the new spectrum to the average spectrum of the calibration set. Although the correlation between individual GH values and prediction errors is weak, Global H is very useful in assessing whether a calibration is appropriate for a new group of samples. As a rule of thumb, an average GH value above 3.0 for a group of samples should be cause for concern.

The Neighbourhood H statistic was developed following the observation that having just a few calibration spectra similar to a new spectrum provided a noticeable improvement in prediction accuracy. The Neighbourhood H statistic uses the same formula as Global H, but measures the distance from a spectrum to the closest calibration spectrum instead of to the average calibration spectrum. This requires that the distance from a spectrum to every calibration spectrum be computed in order to identify the closest calibration spectrum. The best use for NH is during calibration expansion. Not much is gained by adding spectra that are already represented in the calibration set.

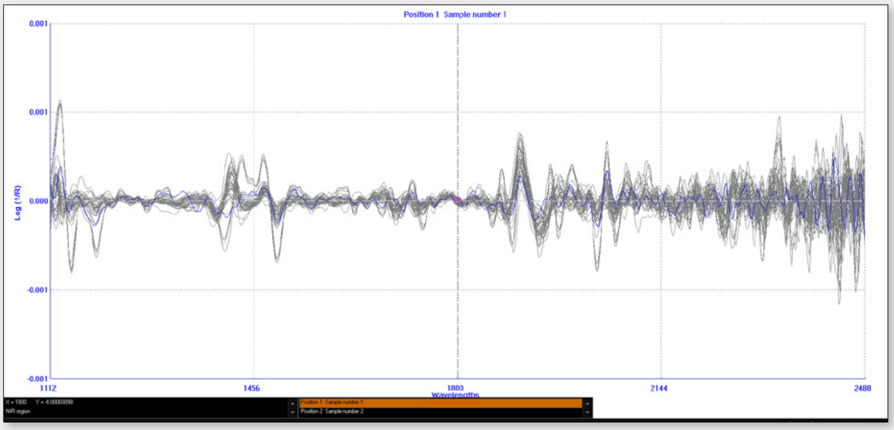

Figure 1. X residuals of milk powder spectra after pre-treating with second derivative and removing 12 PCA loadings.

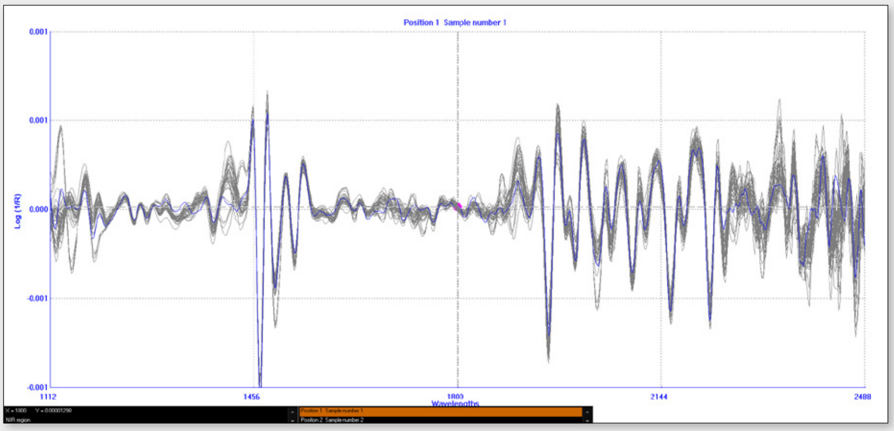

Figure 2. X residuals of milk powder spectra spiked with melamine after pre-treating by second derivative and removing 12 PCA loadings.

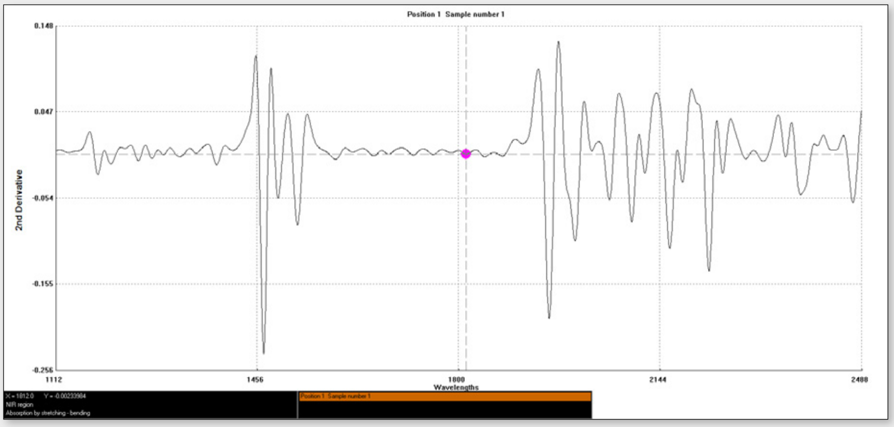

Figure 3. Second derivative spectra of melamine.

The best samples to add to a calibration set are ones that are not well-represented and have a high NH value.

Neighbourhood H values provide additional information during routine analysis. Samples extremely high or low in fat or moisture will likely have high GH values. Reasonable NH values provide assurance that such samples are represented in the calibration and will likely have acceptable prediction accuracy. Global H and Neighbourhood H made a lot of sense when applied to MLR problems because each variable was scaled and treated equally. A problem arose when they were applied to variables selected sequentially via stepwise regression or constructed from loadings via PLS or principal component analysis (PCA). In those cases, the first few variables are more important to the regression than the last few variables. Not only does this give undue influence to the last few variables, but it also gives the number of variables in the model more influence on the GH and NH values.

Good product definition (GPD)

GPD was developed to help monitor a product moving through a pipe. While GH and NH could be used, we wanted a way to protect the product, not the calibration. The GPD program starts with a set of known, good product examples. The usual scatter correction and block derivatives can be applied to the spectra. The program further cleans the dataset by creating a PCA and examining the sample scores. A sample is removed from the set if any of its scores is too large. Once the set is cleaned, the GPD program computes a final PCA. At this point, the number of PCA factors to be used must be picked based on the complexity of the dataset.

Several "distance" measures are available. The first one used was called maximum *X* residual. PCA loadings are removed from a spectrum to obtain the *X* residuals. A typical summary of theses residuals is to take the root mean square of them. Instead, we use the maximum absolute value of the residuals. This measure is more sensitive to deviations limited to a small region of the spectrum. A similar measure is called maximum *X* residual T, in which the standard deviation of the absolute residual is computed at each wavelength in the training set. These standard deviations are used to turn each new residual into a "t" statistic. The largest absolute "t" statistic is then used to represent the spectrum distance.

Another new measure is called maximum peak T. We start with the same *X* residuals as the other distances. This time we check a small region of a spectrum for an unmodelled peak by multiplying the *X* residuals in that region by weights that mimic the shape of an absorption peak. The "T" in the name indicates that we are dividing each weighted sum by the standard deviation of corresponding weighted sums in the training set. Maximum peak T has been shown to be very useful in detecting contaminants in a process.

To demonstrate the ability of maximum peak T to find a contaminant, spectra of milk powder were split into an 81 sample training set, a 40 sample tuning set and a 93 sample testing set based on time splits. The tuning and testing sets were artificially spiked with 0.5% melamine. After pre-treating with a second derivative, a 12-factor PCA was performed on the unspiked training set milk powder samples. The *X* residuals in Figure 1 are the tuning set spectra after removing the 12 PCA factors from the training set. The average is not all zeros, indicating that the tuning set was somehow different from the training set.

The X residuals in Figure 2 are from the same samples as in Figure 1, but contain the 0.5% melamine spiking. Note the additional second derivative peaks.

Comparing the additional peaks in Figure 2 to the second derivative of melamine in Figure 3, it is clear that the additional peaks are due to melamine. The Maximum Peak T statistic finds these additional peaks and flags all spiked samples in the tuning and test sets as being different from the training set. The use of Maximum Peak T for actual (not simulated) contaminant detection is still being evaluated, but appears to allow for a lower detection limit than Maximum *X* Residual or Global H.