Understanding the root cause(s) of nonlinearities in near infrared spectroscopy

Harald Martens

Abstract
NIR process monitoring and NIR hyperspectral video generates a deluge of non-selective spectral data, information-rich but per se useless. This paper demonstrates how interpretable data modelling can lead to simpler and better use of such NIR Big Data: A set of simple powder mixtures of the main constituents in wheat flour were measured by NIR transmission under different measurement conditions. Their absorbance spectra were submitted to multivariate calibration for predicting the protein content, by standard chemometric calibration by PLS regression. A reasonable calibration model was obtained, but it was unexpectedly complex and not robust. However, closer inspection the PLS regression subspace showed a surprising structure. This allowed us to identify the problem: Non-additive, strongly overlapping light scattering and light absorption effects in the NIR absorbance spectra. Based on this insight, a pragmatic, but causal preprocessing model was set up and iteratively optimized for predictive ability. This nonlinear optimized extended signal correction (OEMSC) separated and quantified the main physical and chemical sources of variation in the spectra. The preprocessing greatly simplified the NIR spectra and their quantitative calibration and prediction.

Keywords
NIR, diffuse, multivariate calibration, chemometrics, PLSR, interpretation, EMSC, optimized, OEMSC, wheat protein, powder mixtures, transmittance, explainable AI, causalities, light absorption, light scattering

The body of work produced by Professor Harald Martens on multivariate analysis and near-infrared spectroscopy spans almost 50 years. His contributions to the chemical, biological, food science, agriculture and other industrial fields are important for their insights into both their understanding of the basic science and for providing pragmatic solutions for future applications. Professor Martens is a Fellow of Near Infrared Spectroscopy. He was elected by the International Conference of Near Infrared Spectroscopy in 2015. In Part 1 of this series: Notes on the use of Modern Multichannel Measurements, Harald helps us understand the underlying root causes of nonlinearities of near-infrared spectra using classical chemistry, physics and an innovative chemometric tool developed by Harald and colleagues; Optimized Extended Multiplicative Signal Correction (OEMSC).

Introduction
Now society has an urgent need for tools to let humans understand the information hidden in Big Data. Continuous, high-speed multiwavelength diffuse spectroscopy creates Big Data. Can our field contribute something towards a more humane, robust and efficient artificial intelligence? This author thinks so, based on the way NIR culture has been using interpretable data-driven modelling for several decades, not even knowing that we were using “machine learning”.

Here is an illustration. The light that we see is dominated by two major phenomena: Chemical light absorption is mainly caused by molecules “killing” different amounts of light at different wavelengths, leaving the rest to reach our eyes and cause colour sensation. Physical light scattering is caused by local variations in the refractive index that force the light to change direction. Think of air bubbles in whipped cream, oil droplets in mayonnaise or protein particles in hard-boiled egg.

Gary E. Ritchie, MS, CNIRS News Editor
Corresponding author:
Harald Martens, Prof. Emeritus Big Data Cybernetics, Department of Engineering Cybernetics, Norwegian University of Science & Technology NTNU, 7034 Trondheim, Norway; Chief Research Scientist, Idletechs AS, Havneg. 7, 7010 Trondheim, Norway.
Email: harald.martens@idletechs.com
In fact, most of what we look at is affected by both chemical light absorption and physical light scattering. For humans and animals that is fine: Without light scattering, we would have been transparent. Without light absorption, life would have been grey.

But since chemical light absorption and physical light scattering are very different natural phenomena, their effects on light measurements carry very different types of information. Hence, one given spectrophotometer or hyperspectral camera system can be used for quantifying a wide range of properties. As the readers of this journal very well know, the variations in one single measured spectrum can be converted into a number of different information types, depending on what it has been calibrated for.

Several different chemical compounds may be characterized with respect to their concentration and sometimes even their molecular state. Large groups of naturally correlated chemical compounds may also be quantified. Several different physical variation types may likewise be observed from the light scattering properties, ranging from surface and particle-size and -distribution to simple assessment of sample thickness.

For optimal instrument output, the way the different chemical and physical variations affect the measured spectra need to be modelled mathematically – not perfectly, but with sufficient realism. Some of these causes for spectral variation are known and of interest; others are known, but just a nuisance, and many are simply of unknown type.

In general, all spectral variation types that overlap spectrally with the variations of real interest need to be handled – either kept constant or modelled and thus compensated for mathematically in the multivariate calibration process. To be modelled, their variations must either be observed within the calibration sample set or described based on prior knowledge. Either way, it is advantageous to understand which variations need to be modelled and how.

Multichannel measurements, e.g. from a high-speed scanning spectrophotometer or a hyperspectral video-camera, certainly represent Big Data. But instrument calibration is risky business if performed according to the paradigm “just get yourself some Big Data and apply Artificial Intelligence (AI)/Machine Learning (ML)”. Today, this unfortunate tradition is under heavy attack, both ethically, legally and economically. The new buzzword is “Explainable AI” (XAI).

This paper will show how precise, but highly confusing and non-selective diffuse spectroscopy can be converted into more accurate, precise and understandable information, based on chemometric methods. This will be illustrated using a small set of “on purpose dirty” spectral measurements that Jesper Pram and I made for didactic purposes about 20 years ago at U Copenhagen, in the chemometrics group that prof. Lars Munck originally established there.

The data analytic problems and the solutions are rather generic. Therefore, in case the readers or their colleagues might find them useful also for XAI/ML and Big Data in general, they will mainly be described in non-technical terms.

**Experimental**

The present didactic data set concerns a “classical NIR” topic, protein quantification in dry wheat flour by high-speed NIR transmission spectroscopy. A total of 100 absorbance spectra were collected in the range of 850–1048 nm using an Infratec 1255 Food and Feed Analyzer (FossTecator, Höganas, Sweden) fitted with a standard sample holder for five cylindrical cuvettes. Five different flour mixtures were each measured 20 times under 20 different experimental conditions. For linearization, the measured transmittance $T$ was converted into absorbance $A$ by the usual $A = \log(1/T)$.

More technical information about the experiment is given in Appendixes 1 and 2.

The instrument itself is known to give very precise and accurate measurements. And NIR measurements are fast and simple: Virtually no sample preparation is required – just put the powder into the sample holder with a glass bottom, put on a glass top and measure how much light is transmitted through the sample at the different wavelengths. Virtually no sample preparation, and very simple measurements. That is the charm of NIR.

**Results and discussion**

Figure 1 shows the 100 measured absorbance spectra obtained. Why do these simple two-component mixtures give so unexpectedly “dirty” spectra? How can they be robustly converted into “clean” information?

At least the “dirty” data make some sense: The absorbance spectra display several different “chemical”
absorbance maxima, as expected for mixtures of constituents known to have peak-shaped absorbance spectra.

The varying measurement conditions appear to affect the spectra strongly. How detrimental are these unknown variations, and how should we deal with them?

With sufficiently informative calibration data, “modern machine learning,” e.g. an artificial neural net or a support vector machine, could probably have extracted the desired chemical information from these input absorbance spectra in the calibration sample set. But it is important to maintain control over the calibration process, for unexpected errors may have occurred. Moreover, the number of calibration samples is very limited (only three chemical compositions).

So we choose a simple, linear calibration model, and estimate the calibration model parameters by a traditional, easily validated chemometric calibration technique, the cross-validated Partial Least Squares Regression (PLSR). Figure 2 shows the resulting prediction ability to be reasonably good, although it is not perfect.

Satisfied with that, we now go to the simplest type of XAI: We want to know HOW the machine learning solution converts the many measured channels X into predictions of protein concentration y. That is shown in Figure 3, in terms of the so-called regression coefficient spectrum \( \mathbf{b} \) that summarize the PLS Regression model. A linear combination of high spectral values around 910 and 1030 nm and low levels around 860 and 990 nm is seen to contribute to the prediction of high levels of y.

Similar model summaries may be obtained from a wide range of multivariate statistical regression or ML tools. This information is much better than nothing. At least we see how different aspects of the measurements are balanced against each other for the prediction, and gross mistakes may sometimes be discovered.

But alas, but it does not give meaningful explanation about WHY the different input variables are combined this way in this application. Without that, we are still in the dark as to how reliable and robust this result it.

To get to a deeper level of understanding, we employ a characteristic aspect of the chosen ML method, the PLSR. Namely the low-dimensional subspace it generates to describe the most relevant variation patterns in the spectra. Subspace regressions (e.g. Principal Component Regression (PCR), etc.) replace the many (K) input variables in X by a small set of “supervariables” or “components” from X (linear combinations, after mean-centering and scaling).

The PLSR version of subspace regression searches for orthogonal components that explain maximal X/y covariance. These are then used for modelling y. The best number of subspace dimensions (components) is often determined by so-called cross-validation within the calibration set. Then the model is applied to new samples to predict y from X.

Models with 1, 2 or 3 components here explained 68, 95 and 98% of the cross-validated variance in y in the calibration set, respectively. In the independent test set, with very different y-levels, they explained 18, 88 and 97% variance in y, respectively. So in both cases, the three component-solution gave the “best” predictive performance.

However, this “best” solution has two troublesome aspects:

a. Given that the present samples were just mixtures of two constituents whose sum was 1, we expected to need only one component. But the 1-component solution explained only 18% of the Y-variance in the test set. So why do we need three components?

b. The third component picked only up a very small spectral variation pattern (0.005% of the variance in X). So while it appeared necessary for predicting
y, the three-component model is probably not robust—it may be very sensitive to increased measurement noise. But the two- or one-component models simply are not good enough. Que faire?

**Understandable machine learning with an eye for causalities**

If we could find out what it is that plays havoc in our NIR absorbance spectra, then we could model it explicitly and thereby hopefully develop a more robust calibration model from the same data.

Subspace regression methods give us a unique advantage: We can inspect the first few subspace dimensions graphically, getting insight into what is going on in the data. Perhaps we can thereby get ideas about what has caused the variations, and how they can be modelled more efficiently. And then we can test these ideas.

Inspecting the first two PLSR components (Figure 4, linear combinations of the spectra in Figure 1), we see that the five mixture samples showed an intriguing variation pattern: The different measurement conditions within each of the three calibration mixtures (blue) and two test mixtures (red) form distinct lines, one for each protein percentage tested. Each line ranges from low (top, right) to high powder compression. Their extrapolations meet in a featureless offset spectrum.

Armed with this new spectral insight, we now build a pragmatic, semi-causal preprocessing model and fit it to each spectrum, by a method is called Extended Multiplicative Signal Correction EMSC. The EMSC allows a simultaneous correction for both additive and multiplicative interferences. The additive correction can, e.g. reduce unwanted chemical absorbance interferences according to Beer’s Law, as well as lamp intensity variations and other baseline variation, while the latter correction can reduce unwanted variations in optical path length according to Lambert’s Law.

Building in chemical background knowledge, the EMSC has been applied to these present data. We recently used the EMSC as preprocessing of very large amounts of hyperspectral video data in the VNIR and SWIR ranges, respectively, to monitor the drying of wood in terms of known and unknown chemical and physical variation types.

In order to make the EMSC model easier to set up in practice, the author modified it into the present Optimized EMSC (OEMSC). A new, abstract dimension was added to the EMSC model and fitted iteratively to a given set of calibration data by non-linear simplex optimization, minimizing the prediction error in the calibration samples’ y-data. Once optimized, the OEMSC model is then applied also to test samples etc.

Figure 5 shows that the predictive ability improved after the OEMSC pre-treatment. Moreover, with after pre-treatment the PLSR calibration model only

---

*Figure 4. TRYING to see WHY: Searching for clues about the spectral variations in Figure 1. The score space of the first two PLSR components around the model center, with their spectral definitions.*
required one single PLS component, as expected for these mixtures of two powders.

Figure 6 shows how the PLSR combined the preprocessed absorbance from the different wavelength channels to predict the protein content, in terms of the regression coefficient spectrum $b$ estimated using 1 PLS component. It is quite different from that in Figure 3. This illustrates that the model summary $b$ like the one in Figure 3 may be difficult to interpret spectroscopically.

For the results in Figure 5 and 6, the new, abstract EMSC model dimension has been optimized to give preprocessed spectra with optimal $y$-prediction after it has been subtracted.

Why had the predictive ability been improved and the model complexity been reduced by the OEMSC preprocessing? Figure 7 (left side) shows what this version of the OEMSC preprocessing had done to the 100 spectra in Figure 1.

The effect of light scattering appears to have been eliminated, leaving only absorbance information in the preprocessed NIR spectra. The simplified regression coefficient $b$ in Figure 6 just reveals how the spectra change when the protein fraction increases.

However, the OEMSC can be used in different ways for the same data. In the right subplot the OEMSC has instead been optimized to give preprocessed spectra with optimal $y$-prediction without being subtracted (Figure 7, right side). Method details are explained in the Appendix.

Again, a single-component PLSR model gave best prediction, and the predictive improvement (not shown here) was very similar to those in Figure 5. But now a systematic variation in light scattering is retained in the preprocessed data. The difference in appearance between the two OEMSC solutions in Figure 7 indicates that the two causal constituents, wheat protein and wheat starch powders, differed both in chemical absorbance signature and in physical light scattering. Hence, in this case, both the chemical absorbance and the physical light scattering could in principle be used in a final calibration model.

Is the beneficial effect of the OEMC only relevant for such simple calibration examples? That remains to be seen, but here is the author’s assessment:

The EMSC model can be extended to successfully handle quite complex variation types. The OEMSC just extends the EMSC model with one more dimension and should therefore also be able to work for more complex systems. On the other hand, this small example concerned diffuse transmittance measurements. The lessons learned may be expected to generalize also for diffuse transfectance. But for diffuse reflectance the OEMSC may have to be modified to accommodate the effects of specular reflection and very short optical path length there.

In summary – diffuse multichannel spectroscopy of intact samples can give confusing data, and lots of them. Multivariate calibration by some sort of data-driven modelling can reduce some of the confusion. If a subspace-type regression is used in this “machine learning,” e.g. the PLSR, then a graphical inspecting of the regression subspace may allow us to discover why the spectra vary more than anticipated. This new insight can help us formulate better spectral pre-treatment models. In the present case, this understanding allowed us to formulate a mathematical description of the combined light absorbance/light scattering problems, and to use this in a preprocessing that improve the analyte quantification prediction.

In that sense, the subspace modelling, as practiced in chemometric analysis of diffuse NIR spectra, offers generic tools for hybrid data-and-theory driven modelling. These can be used for people to discover causes of
wanted and unwanted patterns in e.g. quantitative Big Data, within XAI.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

References
1. Martens H, Pram Nielsen J and Balling Engelsen S. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near infrared transmission analysis of powder mixtures. *Anal. Chem* 2003; 75: 394–404.
2. Wold S, Martens H and Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In: Proc. Conf. matrix pencils (A. Ruhe and B. Kågström, eds.), March 1982, Lecture Notes in Mathematics. Heidelberg: Springer Verlag, 2003, pp.286–293.
3. Martens H and Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J. Pharmaceut Biomed Anal* 1991; 9: 625–635.
4. Stefansson P, Fortuna J, Rahmati H, et al. Hyperspectral time series analysis: hyperspectral image data streams interpreted by modeling known and unknown variations. In: Amigo JM (ed.) *Hyperspectral imaging* Vol. 32, 1st ed. Amsterdam: Elsevier 2020, pp.305–331.
5. Vitale R, Stefansson P, Marini F, et al. Fast analysis, processing and modeling of hyperspectral videos: challenges and possible solutions. In: Comprehensive chemometrics. 2nd ed. Amsterdam: Elsevier, 2020 pp.395–409.
6. Martens H. The informative converse paradox: windows into the unknown. *Chemometr Intell Lab Syst* 2011; 107: 124–138.

Appendix 1. Modelling details.
In both OEMSC cases, the basic EMSC model consisted of the average of the calibration spectra in Figure 1 (yielding parameters describing the scaling of the path length, relative to the average path length), and a flat “baseline spectrum” [1,1,1,1,...,1] and a linearly sloped “baseline spectrum” [1: +1] (yielding additive deviations from the average baseline). For each spectrum, the relative scaling was divided away, while the additive deviations were subtracted.

The present OEMSC version builds on an implementation originally made by Morten Beck Rye in his MSc thesis based on the author’s proposal. An extra additive spectrum was estimated as a linear combination of the five first principal components loadings of the input spectra in Figure 1, with five unknown parameters estimated by minimizing the prediction error in y. In Figure 7 (left), it was optimized for its effect to subtracted in the preprocessed spectra along with other additive EMSC effects. From Figure 5 (right), it was optimized for its effect being retained in the preprocessed spectra.
Appendix 2. Modelling details

For each individual spectrum \( z_i \) a linear EMSC model is defined

\[
Z_i = c_{i,1} * z_{\text{cal}} + c_{i,2} * z_{\text{flatBaseline}} + c_{i,3} * z_{\text{slopedBaseline}} + c_{i,4} * z_{\text{calOEMSC}} + \epsilon_i
\]

(1)

where \( \bar{z}_{\text{cal}} \) is the average of the calibration absorbance spectra (here: 60 blue curves in Figure 1), \( z_{\text{flat}} = [1, 1, 1, \ldots, 1] \) and \( z_{\text{slopedBaseline}} = \text{e.g.} [-1, \ldots, 0, \ldots, +1] \). The unknown component spectrum \( z_{\text{calOEMSC}} \) represents a spectral variation type – expected or unexpected – that is so dominant that it clearly disrupts the calibration unless included in the EMSC model. How to estimate it will be described below.

The OSEM score \( c_{i,1} \) describes the scaling, e.g. the effective optical path length relative to the average path length. Scores \( c_{i,2} \) and \( c_{i,3} \) represent additive deviations from the average baseline, while \( c_{i,4} \) represents the amount of the unknown component that, when optimized, makes the EMSC model more physically realistic and thereby improving the calibration.

For each input spectrum \( z_i \), its unknown scores \( c_{i,1}, c_{i,2}, c_{i,3} \) and \( c_{i,4} \) are estimated by minimizing the sum of squares in the residual spectrum \( \epsilon_i \).

The resulting estimated residual spectrum \( \hat{\epsilon}_i \) is expected to represent most of the uninteresting measurement noise and modelling errors. But it can also contain a systematic spectral variation that was too small to disrupt the modelling, but was the reason for measuring the NIR spectra in the first place.

The actual preprocessing of \( z_i \) then consists of dividing away the effect of the scaling score in order to adjust all spectra to the same optical path length, while the additive deviation scores are subtracted in order to adjust them to the same common baseline.

If the unknown component is considered useless, then the preprocessing of absorbance spectrum \( z_i \) is

\[
x_i = (z_i - \hat{c}_{i,2} * z_{\text{flatBaseline}} - \hat{c}_{i,3} * z_{\text{slopedBaseline}} - \hat{c}_{i,4} * z_{\text{calOEMSC}}) / \hat{c}_{i,1}
\]

(2a)

or equivalently, a normalization of the remaining spectral information

\[
x_i = \hat{\epsilon}_i / \hat{c}_{i,1}
\]

(2b)

If instead the unknown component is considered potentially useful and therefore retained, the preprocessing is simply

\[
x_i = (z_i - \hat{c}_{i,2} * z_{\text{flatBaseline}} - \hat{c}_{i,3} * z_{\text{slopedBaseline}}) / \hat{c}_{i,1}
\]

(3a)

or equivalently, showing that the unknown variation is retained in the preprocessed spectra \( x_i \),

\[
x_i = (\hat{\epsilon}_i + \hat{c}_{i,4} * z_{\text{calOEMSC}}) / \hat{c}_{i,1}
\]

(3b)

What characterizes the optimization step in the OEMSC is that initially unknown difference spectrum \( z_{\text{calOEMSC}} \) is defined as a weighted sum of the loading eigenvectors of the first few (e.g. 5) principal components of the input calibration spectra in Figure 1

\[
z_{\text{calOEMSC}} = t_1 p_{1}^t + t_2 p_{2}^t + t_3 p_{4}^t + t_4 p_{4}^t + t_5 p_{5}^t
\]

(4)

The motivation for this is that the unknown, but dominant variation patterns, which \( z_{\text{calOEMSC}} \) is intended to pick up and thereby protect the rest of the EMSC modelling from, is – almost per definition – expected to be evident inside the low-dimensional space of the first few principal components of the set of calibration spectra \( Z = [z_i, i = 1, 2, \ldots] \). To find the “best” \( z_{\text{calOEMSC}} \), the unknown weights \( t_1, \ldots, t_5 \) are simply found optimizing a chosen criterion, e.g. by standard nonlinear simplex optimization. Here, we minimized the sum of squared prediction errors in the linear model \( \Sigma(y - \hat{b} - X * b)^t \) where \( y \) is the protein fraction and \( X = [x_1, x_2, \ldots] \) is the preprocessed spectra in the calibration sample set.

In Figure 7 (left), \( z_{\text{calOEMSC}} \) was optimized for its effect to subtracted (equation (2a)). Figure 7 (right), it was optimized for its effect being retained (equation (3a)).

It should be noted that the OSEM scores \( \hat{c}_{i,1}, \hat{c}_{i,2}, \hat{c}_{i,3}, \hat{c}_{i,4} \) do not necessarily represent mistakes – they can be highly informative in their own right. And contrary to many other preprocessing methods (smoothing, etc.), the OSEM preprocessing is lossless, in the sense that the input spectra \( z_i \) (Figure 1) can always be reconstructed from the output spectra \( z_i \) (Figure 7) and the corresponding OSEM scores \( \hat{c}_{i,1}, \hat{c}_{i,2}, \hat{c}_{i,3}, \hat{c}_{i,4} \) and the estimate of the unknown spectrum \( z_{\text{calOEMSC}} \).