The wonderful world of visible-near infrared spectra: theory and practice

John S. Shenk

The Pennsylvania State University and Infrasoft International, LLC, 109 Sellers Lane, Port Matilda, PA 16870, USA

Introduction

When did this technology begin? It began 15 billion years ago with the formation of electro-magnetic radiation. Electro magnetism is one of the four basic forces of the universe. The other three are gravity and the strong and weak forces inside the atom. Cosmologists tell us that the electro-magnetic force became present in the universe 350,000 years after the big bang. My area of interest has always been biology. Biology is heavily dependent on the visible (vis) region of electro-magnetic radiation to supply the energy to support life on this planet. Matter interacts with visible radiation by absorbing energy that causes electrons to jump to a higher energy level. The near infrared (NIR) spectrum is the region next to the visible with longer wavelengths. It contains information about the functional groups CH, NH and OH contained in all biological substances. Matter interacts with NIR radiation by absorbing energy that corresponds to the frequency of the bending and stretching vibrations in chemical bonds. The region from 800 nm to 1350 nm contains both sources of information.

Chemometrics—the tools of our trade

Chemometrics is a discipline concerned with the application of statistics and mathematics to chemistry. The vis-NIR spectrum obtained by reflectance, transmission, or folded transmission is a composite of all the chemical and physical information of the sample that interacts with the radiation. To best utilise this information for practical applications, chemometric techniques are needed to turn the spectrum into estimates of sample composition. A very brief table of these tools would include the following procedures:

Table 1. Chemometric tools.

1. Derivatives—(help to remove spectral redundancy)
2. Scatter corrections—(help to remove scatter and non linearity)
3. Scores—loadings (PCA, PLS) (product libraries) (find the right samples)
4. Regression methods—MLR, PLS and all possible modifications, neural networks etc., discrimination—(mathematical prediction models)
5. Instrument spectral matching—(correct spectra to a master instrument)
6. Monitor—(statistical comparisons among spectra) (actual vs predicted)
7. Graphics—displays

The radiation as applied by most instruments has very little energy and penetrates only a millimetre or so into the substance depending on the substance's surface composition and structure. By exposing the surface of a plant leaf to vis-NIR radiation, the energy is either reflected like a mirror, or scattered within the leaf where it is either absorbed by the leaf's electrons or chemical bonds or re-emitted. This very complex spectrum is a function of the physical characteristics and

Near Infrared Spectroscopy: Proceedings of the 11th International Conference.

chemical composition of the substance. The mathematics behind the chemometrics procedures makes this complex spectrum useful in hundreds of practical applications.

When radiation is scattered, it gathers information about the absorption pattern of the sample, but less of the radiation finds its way to the instrument's detectors, making the absorptions appear even stronger. Scatter stretches the composite absorption curve upward. If the material being evaluated is ground before collecting the scan, the small size of the particles of the ground substance will reduce the scatter effect. The larger the particle, the greater will be the stretching of the absorption peaks.

Surface reflectance has the inverse effect of scatter. If a sample surface is shiny, some of the radiation will be reflected like a mirror without the opportunity to interact with the sample, lowering the absorption peaks. If some of the reflected light reaches the detectors, the composite curve is further squashed downward.

Every substance has a unique vis-NIR composite spectrum. It consists in its simplest form as a combination of radiation scatter, surface reflectance, and absorption of chemical bonds. If two samples of a material have the exact same spectrum, we can assume that they have nearly the same physical (scatter and surface reflectance) and chemical composition. The more different the spectra, the larger will be the physical or chemical sample differences. The vis-NIR spectrum is unique for each and every biological substance. Quantification of this information is the next challenge in making vis-NIR a practical analytical tool.

Chemometric patterns in multiple dimensions—selecting the right samples

The NIR spectrum is a representation of the total chemical and physical composition of a material. Although the information is unique for each material or sample of the material, with a few exceptions at present, it cannot be used directly. The primary reason is that the current analytical system is based on chemical extractions of pure or semi-pure substances in a laboratory environment. This relationship of vis-NIR spectra to the current reference method procedures will be discussed in a later section of this paper. Here in this section we want to better understand how the vis-NIR measurement is used in the analytical system.

If we take a group of spectra from samples of a given material, we can characterize the two dimension spectra in multiple dimensional space by a loading-score method, Figure 1. The loading-score concept converts the two dimensional wavelength space into a multiple dimensional sample space, where each sample is a point. The loadings represent the independent patterns in the set of data and the scores represent the proportion of each pattern in each sample's spectrum. This conversion from wavelength space to multiple dimensional space has two main advantages. First, each pattern is calculated to be independent, removing the redundancy of information within each spectrum. Second, since each sample is represented by a point, a distance between samples can be computed. We measure the distance between a sample and the location of the average sample. We also measure the distance between a sample and the sample closest to it. We refer to these measurements as the global H (H = Mahalanobis distance), and neighbourhood H.

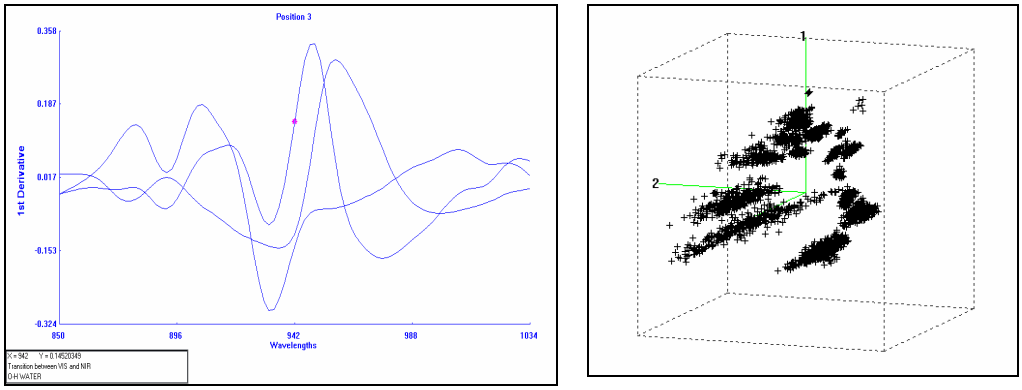


Figure 1. A display of the first three loadings and a 3D display of the product library.

Chemometric 3D displays of the product libraries

By transforming the spectra to loading-score mathematics, we have gained another important feature not found in most other analytical techniques. This is the capacity to identify spectra unlike any spectra in the current library of spectra representing the product. This feature is important so that these unique samples can be identified and inspected or analyzed by other means. A second feature is that if a sample is analyzed in routine analysis and it does not have an acceptable GH or NH, the operator can save the sample and its spectra to help expand the product library representing the material, Figure 2.

We now have a better understanding of the NIR spectrum and how to group samples into product libraries, and a simple test to use in routine analysis to identify new samples that should be added to the product libraries. To make practical use of the spectrum we must relate it to something we know. This known information is usually obtained in a chemistry laboratory and referred to as reference analysis.

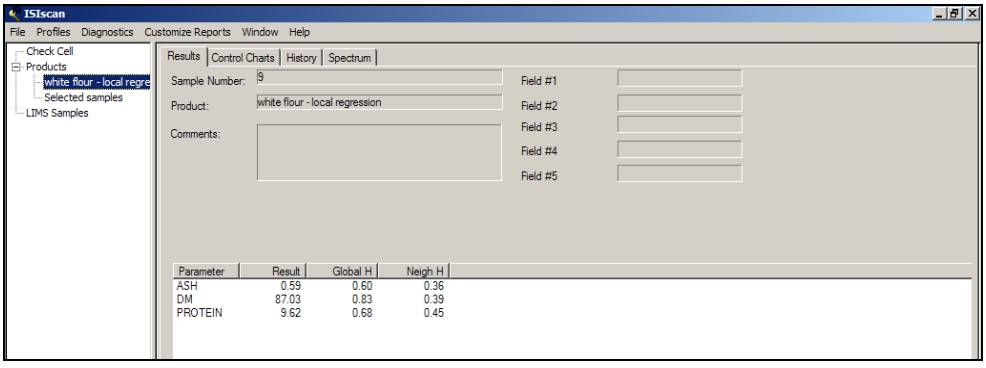


Figure 2. The GH–NH values in routine analysis.

The most popular way to use this analytical technology is to calibrate it with reference values obtained from a laboratory or some other source. As stated above, interpretation of the absorption information is extremely complex and must be related to some known analytical value to be of practical importance. Typically most agriculture and food applications use NIR to predict protein,

fat, fibre, and moisture in a material. The quantification is accomplished by obtaining reference values from a laboratory for all samples in the product library.

A number of regression methods are available to make the prediction equation or model. Popular methods are stepwise regression, partial least squares regression,[1,2] and neural networks.[3] If all samples are used in the library, the calibration equation is referred to as a global calibration. Global calibration equations can be developed either as linear models through regression or non-linear models through procedures like neural networks. In addition, new procedures are now available to develop LOCAL calibrations[4] for each sample-constituent of the product during routine operation, Figure 3. The most important part of the calibration process is not choosing the regression technique, but choosing the right samples for the product library development.

WF11	PROTEIN	10.40	10.59	-0.19	1.14	0.16	200
WF11	ASH	0.82	0.84	-0.02	1.98	0.23	200
WF11	DM	87.13	87.17	-0.04	1.22	0.21	200
WF12	PROTEIN	9.92	10.23	-0.31	0.58	0.14	200
WF12	ASH	0.54	0.54	-0.01	0.81	0.12	200
WF12	DM	86.90	86.68	0.22	0.59	0.12	200
WF13	PROTEIN	0.00	12.52	0.00	1.16	0.34	200
WF13	ASH	0.00	0.58	0.00	0.45	0.27	200
WF13	DM	0.00	86.13	0.00	0.63	0.26	200
WF14	PROTEIN	13.00	13.18	-0.18	1.03	0.19	200
WF14	ASH	0.62	0.60	0.02	0.46	0.14	200
WF14	DM	85.90	86.03	-0.13	0.33	0.15	200

<> Summary

Variable	NumPred	Total	SEP	Bias	SEP(C)	Slope	RSQ	Ave GH	Ave NH
PROTEIN	14	4445	0.258	-0.117	0.240	0.952	0.976	1.000	0.365
ASH	14	4445	0.018	-0.011	0.015	1.002	0.987	0.969	0.292
DM	14	4445	0.217	0.043	0.221	1.001	0.691	0.761	0.266

Figure 3. A display of LOCAL analysis

Chemometrics to tell us how well we are doing

Two important sources of error must be considered to evaluate the accuracy and repeatability of our predictions. Accuracy is the agreement between the reference value and the predicted value from the NIR spectrum. Repeatability is the agreement among sub-samples analysed from the same material, or sub-samples analysed by different instruments. The errors of accuracy and repeatability determine the usefulness of the analysis. In general repeatability is almost always very acceptable, but the accuracy is a function of the material, constituents within that material, and the repeatability of the laboratory procedure developing the reference values, Figure 4. To give some insight into what has and is being analysed by this technology, let's look at the applications.

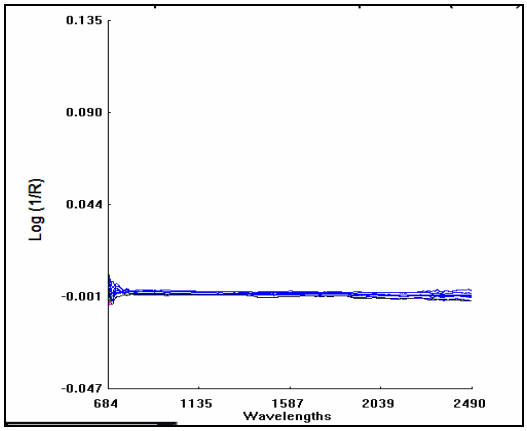
SEP:	0.372	Number of Samples:	1840
Means:	13.765 13.649	Standard Deviations:	2.452 2.475
Bias:	0.116	Bias Limit:	0.081
SEP(C):	0.354	SEP(C) Limit:	0.175
Slope:	0.981	RSQ:	0.980
Ave. Global H:	1.330	Ave. Neighbor. H:	0.708

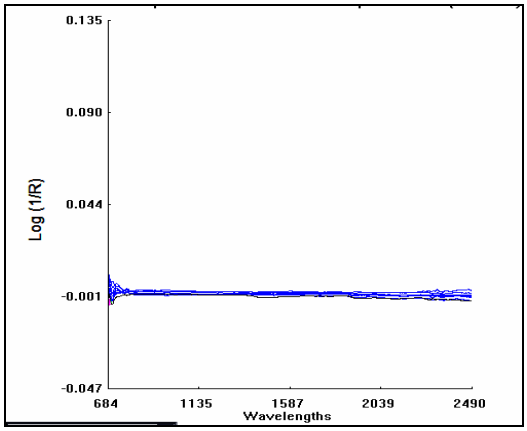
Pos.	Sample #	LAB	ANL	Residual	Bias	Global H	Neigh. H
65	1007	14.877	14.672	0.205	0.088	0.595	0.000
66	1214	17.128	16.918	0.210	0.094	0.914	0.413
67	1835	14.828	14.930	-0.102	-0.219	1.014	0.665
68	216	14.720	14.730	-0.010	-0.126	0.641	0.316
69	2038	13.463	12.913	0.550	0.434	0.681	0.412
70	2056	14.662	14.308	0.353	0.237	1.096	0.712
71	2177	16.624	16.355	0.268	0.152	0.686	0.286

Figure 4. A comparison of actual vs predicted values.

Chemometrics to make the spectrum from instruments alike

Today many agriculture and food companies have multiple instruments. We call the chemometric procedure that tries to make instrument more alike, instrument standardization.[5] One of the most important goals to achieve with multiple instruments is to have the spectrum of a sample scanned on each host instrument the same as the spectrum obtained on a master instrument. When this is accomplished, all instruments in the group produce the same spectrum, the same predicted values and the same GH and NH tests. The goal is to have the repeatability error of the same sample scanned and predicted on different instruments to be of the same magnitude as the error of predicting sub-samples of the same material scanned and predicted on one instrument. There are many mathematical techniques proposed to accomplish this goal. Our method of single sample standardization works well across the Foss family of instruments, Figures 5 and 6.





Figures 5 and 6. Difference spectra between instruments before and after standardization.

Applications

The NIR technology briefly described above is used daily in a vast number of applications around the world. In agriculture the technology provides nutritional analysis of feeding rations for millions of dairy cows. The feed industry serving the poultry and swine industry is using NIR reflectance broadly as a reliable analytical method. The analysis of grains by NIR transmission serves as the marketing tool for feed and food supplies. In the food industry the large confectionery companies use NIR to measure their raw materials and finished products. The technology is used by the dairy industry in quality control of butter and cheese. And the meat industry surrounding the fast food business relied heavily on NIR to control the quality of their products. The list goes on and on. Eventually the technique will be available from the Internet.[6] Near infrared analysis is no longer an analytical procedure to be used by universities and research institutions. It is the only analytical procedure that can possibly produce rapid and accurate information for our modern electronic analytical system**.**

References

1. J.M. Brenchley, U Horchner and J.H. Kalivas. *Appl. Spectrosc*. **51(5),** 689 (1997).
2. M. Forina, C. Casolino and P Millan. *J. Chemometr*. **13,** 165 (1999).

3. N.B. Buchmann and I.A. Cowe, “Advantages of using artificial neural networks techniques for agriculture data”, in *Near Infrared Spectroscopy: Proceedings of the 10th International Conference*, pp. 71–75 (2002).
4. J.S. Shenk and M.O. Westerhaus. (Local) US Patent No. 5,798,526 August 25, 1998.
5. J.S. Shenk and M.O. Westerhaus. (Standardization) US Patent No. 4,866,644 September 12, 1989.
6. J.S. Shenk and M.O. Westerhaus. (RINA) US patent pending serial number 09/633,891. Symbols