NIRS Vision 4.1

Vision

Manual - Theory

8.105.8051EN / 2015-12-04

Metrohm

Metrohm AG
CH-9100 Herisau
Switzerland
Phone +41 71 353 85 85
Fax +41 71 353 89 01
info@metrohm.com
www.metrohm.com

NIRS Vision

Manual – Theory

8.105.8051EN / 2015-12-04 ahag

Teachware
Metrohm AG
CH-9100 Herisau
teachware@metrohm.com

Although all the information given in this documentation has been checked with great care, errors cannot be entirely excluded. Should you notice any mistakes please send us your comments using the address given above.

Table of contents

1 Calibration Development

1.1 Simple Linear Regression

1.1.1 Overview

The simplest method of calibration is based on a single independent variable (wavelength). Vision uses the Inverted Least Squares (ILS) method. The assumption is that constituent values ***c*** (usually concentrations) are a linear function of the absorbance ***A*** at some wave-length ***i***:

$$c = K(0) + K(1) \cdot A_i$$

(This expression contrasts the Lambert-Beer Law, which gives absorbance as a function of concentration, and hence is referred to as "inverted.")

ILS is presented graphically by a least squares fit of a straight line to data points in plot of concentration versus absorbance. The least squares approach fits the line so as to minimize the sum of squares of deviations between data points and the calibration line; it yields two calibration constants: intercept ***K***(0) and slope ***K***(1). Those deviations are taken along the concentration axis, and so are called concentration residuals.

1.1.2 Correlation

The wavelength at which the calibration is performed is usually selected because of high **correlation** between concentration and absorbance: correlation is a measure of how well the data at a certain wavelength is represented by the linear equation. Correlation can range between -1 and 1, with values at either extreme indicating that the linear model does indeed describe the data set (there is a good fit between the concentration-absorbance data and the linear equation).

Often a high absolute value for correlation at some wavelength confirms what is visually evident when sample spectra in a calibration set are overlaid. Such visual evaluation of spectral data is important to maximize confidence that the model developed is indeed appropriate. For, though statistics are important and useful tools, they should be used to help make optimal choices rather than blind choices.

A value for correlation falling in the middle of the range (correlation is close to zero) indicates lack of relationship between constituent and spectral data at this wavelength. In other words, the linear model does not describe the data set. Vision contains tools to select the calibration wavelength that gives the best correlation.

1.1.3 Sensitivity

Another important property of a calibration is **sensitivity**. Sensitivity is the absolute value of the slope term in the calibration equation. If the sensitivity is very high, even small changes in absorbance, e.g. those caused by noise, will register as a large change in predicted concentration. On the other hand, low sensitivity indicates a robust calibration that gives stable and reliable predictions.

Of course, low sensitivity is useful only when correlation is very high, for low sensitivity with low correlation would have little predictive value. Superimposed plots of correlation and sensitivity values across the full spectral range are helpful in visualizing which wavelength(s) is (are) best for calibration. Thus, that wavelength is best which has a correlation close to 1 or -1, as well as low sensitivity.

1.2 Multilinear Regression

1.2.1 Overview

Multilinear Regression (MLR) is an extension of the simple linear regression. This method uses information at more than one wavelength to create a calibration equation of the following form:

$$c = K(0) + \sum_{i=1}^{n} \left[K(i) \cdot A_i \right]$$

MLR is useful when the information at a single wavelength does not yield a model that performs suitably well.

1.2.2 Colinearity

When using more than one wavelength in MLR, there is a risk of colinearity among the chosen wavelengths. This means simply that absorbance values at two or more of the wavelengths used for the calibration are correlated: they describe behavior in the data set that is related, not unique.

While such a model may describe the calibration set very well, it may be very sensitive to noise or systematic errors in the calibration data which may not representative of samples in general. Consequently, MLR models with colinearity among analysis wavelengths may not provide reliable analysis of real samples. Vision provides an intercorrelation table for evaluating the extent of colinearity in an MLR model.

1.2.3 Multiple Correlation Coefficient

A multi-wavelength analog of the correlation coefficient is the multiple correlation coefficient. This parameter has the same significance as simple correlation coefficient, and is similarly useful for estimating the quality of a model. As for simple correlation, the multiple correlation coefficient can have values ranging from zero (0) to one (1), with zero indicating a complete lack of relationship between spectral and constituent data, and one (1) signifying a perfect fit.

1.3 Partial Least Squares (PLS)

1.3.1 Overview

Partial Least Squares (PLS) is a regression method that allows use of many wavelengths – whether broad segments or even the entire spectrum – while avoiding the problem of colinearity that besets MLR. Unlike traditional least squares methods, PLS does not assume that spectral data are exact and all errors are in constituent values. Instead, the spectral and constituent data are simultaneously modeled in steps that incrementally account for spectral signal and constituent values. In each step, part of the spectral data (called a "factor") and a corresponding part of the constituent data is subtracted from the data set, leaving spectral and constituent residuals.

With the determination of each factor, the residual information in the calibration data set (the information yet to be modeled) becomes smaller and smaller. Partial calibrations for each factor (**loadings** for spectral data and **scores** for constituent values) are used to calculate the amount of variance modeled for each factor. At the end of the process, they are assembled into one overall calibration equation.

1.3.2 Preprocessing Calibration Data

Spectral and constituent data are routinely preprocessed before the PLS calibration. (Here, the term "preprocessing" refers to calculations performed during the PLS calibration, not operations on spectra before calibration such as taking the second derivative.) The vector of constituent values is mean centered and scaled to variance of one (1). The set of training spectra is always mean centered. Of course, the mean values and scaling factors are accounted for in the final calibration equation.

Before calculating spectral loadings, another calculation is performed to determine how effectively the data at each wavelength explains residual concentration values. (It is analogous to the correlation coefficient described for MLR.) The result is called a **weighting vector**, or simply a **weight**. Vision scales the data so weights are proportional to the product of correlation and variance in the spectral data, with the result that wavelengths with high absorptivity are emphasized.

1.3.3 Selecting the Number of Factors

Generally, more PLS factors can be calculated than are appropriate for use in a final calibration. Deciding how many factors to use is an important part of PLS calibration. With too few factors, the calibration accounts for too little information and gives correspondingly high prediction errors. When too many factors are used, the model overfits the calibration data (noise or systematic errors unique to the training is included in the model), resulting in a model that is not robust or stable.

Usually the optimal number of factor is established during cross validation or using the external prediction set. When a validation set is available, Vision calculates the **PRESS** (Prediction Residual Error – Sum of Squares) for each factor, then recommends use of the factor having the minimum PRESS value.

1.3.4 Cross Validation

As an alternative to using validation samples in the calculation of PRESS, a cross validation can be performed using the training set. In cross validation, samples in the training set are grouped into subsets. Such a subset may contain several, or only one sample.

During cross validation, one subset is withheld while a calibration is created with the remaining training samples. Then, the resulting calibration is used to analyze samples in the subset as unknowns. Finally, the predicted constituent values are subtracted from the reference (lab) values, and their differences squared and summed. The first subset is returned to the training set, and in turn every remaining subset is analyzed in the same fashion as the first. The resulting PRESS value at each factor is an indicator of how well PLS model performs. A related indicator of performance is MSECV (Mean Squared Error of Cross Validation).

1.4 Statistical Evaluation of Calibration Equations

Several parameters calculated during calibration or prediction of the validation set indicate the quality of the calibration equation its usefulness in predicting unknowns.

1.4.1 Multiple Correlation Coefficient

Multiple Correlation Coefficient (R^2) is a measure of how well the spectral data fit the constituent values. This statistical quantity, also called Coefficient of Multiple Determination, is equal to zero (0) when spectral response is unrelated to constituent data (the relationship is statistically random). A value of one (1) signifies that the constituent values fit spectral data perfectly and all residuals are

equal to zero (0). Formally R^2 indicates the fraction of total variance in the data set modeled by the equation.

1.4.2 Standard Error of Calibration

Standard Error of Calibration (SEC) is a statistical parameter that indicates the upper limit of accuracy in future predictions. When the calibration equation is applied to the training set itself, the SEC is calculated from residuals ***f*** as:

$$SEC = \sqrt{\frac{\sum f_i^2}{N - K - 1}}$$

where ***N*** is the number of samples and ***K*** number of wavelengths or factors.

1.4.3 Standard Error of Prediction

The calculation for Standard Error of Prediction (SEP) is similar in form to that for SEC, except that ***f*** now denotes residuals obtained from the prediction of the samples not used in the calibration, i.e. a validation sample set. Also, the denominator of the expression does not include K. Usually SEP is larger than SEC.

1.4.4 The F-Statistic

F value (or the F-test statistic) is defined is

$$F = \frac{R^2(N - K - 1)}{K(1 - R^2)}$$

where ***N*** is the number of samples, ***K*** is the number of wavelengths or factors, and $\boldsymbol{R^2}$ is the multiple correlation coefficient. The F value is a useful estimate of goodness of fit of spectral and constituent data. It can be also used as a tool for evaluation of how many wavelengths or factors should be used in an equation, and for determining which samples to eliminate as outliers from the calibration set.

1.4.5 Bias

Bias is the average value of residuals calculated from constituent values of a prediction set. Bias value close to zero (0) indicates that the deviations are distributed randomly. A bias value (either positive or negative) that is large compared with typical constituent values indicates systematic error, e.g. changes in the instrument, the condition of samples or the system being analyzed, or in the reference analysis. In some cases, simple bias correction can be used to address this problem.

2 Math Pretreatment Methods

2.1 N-Point Smooth

This is a boxcar type of smoothing. The method's parameter, segment size, defines the size of the boxcar in nanometers. The average spectral value over the segment is placed in the middle in the segment.

The actual number of data points ***s*** in a segment can be calculated from the following equation:

Where ***x*** is the declared segment size, ***ODD*** is a function that rounds its argument up to the

$$s = ODD\left[INT\left((x+3)/2\right)-1\right]$$

nearest odd integer, and INT is a function that rounds its argument down to the nearest integer.

A consequence of the equation for s is that a smooth value cannot be calculated for a certain number of data points at the beginning and the end of a spectrum. This is so because a number of data points on each side of a central data point are used in the calculation of an average value for the segment. Those data points are not available for model development. The number of points "lost" at each end of the spectrum is given as:

$$\frac{s-1}{2}$$

2.2 First Derivative

The first derivative is used most commonly to eliminate baseline offset variation within a set of spectra. As a constant (zero order) term added to a function ***f(w)*** (the spectrum), offset ***C*** is eliminated by taking the derivative with respect to ***w*** (wavelength):

$$\frac{d}{dw}\left[f(w)+C\right] = f'(w)$$

While a different offset value ***C***; is associated with each spectrum ***f(w)***; all are eliminated in the first derivative.

Spectral offset may vary within a set of spectra for many reasons including:

- Particle size differences among samples
- Varying particulate levels between liquid samples
- Small changes in instrument response due to short term variation in lamp intensity, detector response, or instrument temperature

2.2.1 Calculation of the First Derivative

One way to calculate the first derivative is as the first order finite-difference derivative. This method requires two values to be specified: the length of the **segment** (segment size); and the length of the **gap** between segments (gap size).

The first derivative calculation begins by identifying two segments of the specified size at one end of the spectrum with a gap between them of the specified size. Next, the average absorbance values within the first and second segments (**A** and **B**, respectively) are calculated. Finally, the first derivative

value computed as B-A is assigned to the data point in the middle of the gap. Then the whole segment-gap-segment sequence is shifted one data point and the calculations repeated until a first derivative value has been calculated for all data points in the spectrum.

2.2.2 Segment and Gap Size

$$s = ODD\left[INT\left((x+3)/2\right)-1\right]$$

The actual number of data points ***s*** in a gap and a segment can be calculated from the following equation:

Where ***x*** is the declared size, ***ODD*** is a function that rounds its argument up to the nearest odd integer, and ***INT*** is a function that rounds its argument down to the nearest integer.

A consequence of the equation for ***s*** is that a first derivative value cannot be calculated for a certain number of data points at the beginning and the end of a spectrum. This is so because a certain, minimum number of data points are required to calculate **A** or **B**. The number of points ***m*** that are missing at either end of the spectrum for which the second derivative value cannot be calculated is given as:

$$m = s + \frac{g-1}{2}$$

Where ***s*** and ***g*** are the actual numbers of points per segment and gap, respectively (not the specified size, but the actual number).

If the segments or gap contain an even number of data points, the averages and first derivative value would correspond to midpoints between actual data points. Therefore some restrictions are imposed on the number of points in both segment and gap. The following table gives the actual number of points in segment (s) and gap (g) corresponding to various specified sizes.

Specified Segment Size	Points per segment, *s*
1 – 2	1
3 – 6	3
7 – 10	5
11 – 14	7
15 – 18	9

Specified Gap Size	Points per gap, *g*
1 – 2	1
3 – 6	3
7 – 10	5
11 – 14	7
15 – 18	9

2.3 Second Derivative

The linear baseline of a spectrum is described by the first order equation $a \cdot w + C$ (***a*** is slope, ***w*** is wavelength, and ***C*** is offset), which adds to a function ***f(w)*** (the spectrum). As was shown earlier, calculation of the first derivative with respect to ***w*** eliminates the offset term, ***C***. However, the slope term becomes a constant (zero order) term in the first derivative:

$$\frac{d}{dw}\left[f(w)+\left(a \cdot w + C\right)\right] = f'(w) + a$$

Consequently, if each spectrum ***f(w)***; in a data set exhibits a slightly different slope ***a***; then the first

derivative spectra will exhibit offset variation.

It is common practice, therefore, to take the second derivative with respect to ***w*** (wavelength) so as to eliminate both offset and slope:

$$\frac{d^2}{dw^2}\left[f(w)+(a\cdot w+C)\right]=f''(w)$$

Spectral offset and slope may vary within a set of spectra for any of several reasons including:

- Particle size differences among samples
- Varying particulate levels between liquid samples
- Small changes in instrument response due to short term variation in lamp intensity, detector response, or instrument temperature.

2.3.1 Calculation of the Second Derivative

One way to calculate the second derivative is as the second order finite-difference derivative. This method requires two values to be specified: the length of the **segment** (segment size); and the length of the **gap** between segments (gap size).

The second derivative calculation begins by identifying three segments at one end of the spectrum, each separated from by other by a gap. Average absorbance values are calculated for the first, second, and third segments (**A**, **B** and **C**, respectively). The second derivative value computed as A-2B+C is assigned to the midpoint of the second segment. Then the whole sequence of three segments and two gaps is shifted one data point and the calculations repeated until a second derivative value has been calculated for all data points in the spectrum.

2.3.2 Segment and Gap Size

The actual number of data points s in a gap and a segment can be calculated from the following equation:

$$s=ODD\left[INT\left((x+3)/2\right)-1\right]$$

Where ***x*** is the declared size, ***ODD*** is a function that rounds its argument up to the nearest odd integer, and ***INT*** is a function that rounds its argument down to the nearest integer.

A consequence of the equation for ***s*** is that a second derivative value cannot be calculated for a certain number of data points at the beginning and the end of a spectrum. This is so because a certain, minimum number of data points are required to calculate **A** or **B**. The number of points ***m*** that are missing at either end of the spectrum for which the second derivative value cannot be calculated is given as:

$$m=s+g+\frac{s-1}{2}$$

Where ***s*** and ***g*** are the actual numbers of points per segment and gap, respectively (not the specified size, but the actual number).

If the segments or gap contain an even number of data points, the averages and first derivative value would correspond to midpoints between actual data points. Therefore, the number of points in both

segment and gap are subject to the same restrictions defined for the first derivative:

Specified Segment Size	Points per segment, *s*
1 – 2	1
3 – 6	3
7 – 10	5
11 – 14	7
15 – 18	9
19 – 22	11

Specified Gap Size	Points per gap, *g*
1 – 2	1
3 – 6	3
7 – 10	5
11 – 14	7
15 – 18	9
19 – 22	11

2.4 Third and Fourth Derivative

The third and fourth derivatives are calculated as higher order finite-difference derivatives, analogous to the methods described above for first and second derivatives.

2.5 Standard Normal Variate

Standard Normal Variate is a scatter correction method used commonly to normalize spectra when the effective pathlength varies among samples in a data set. Such pathlength variation can occur when measuring the spectra of granular or powdery samples because a) sample presentation in a cell is not perfectly reproducible, or b) particle size varies between samples.

The spectrum is mean centered and then divided by its standard deviation:

$$s_i^{SNV} = \frac{s_i - \bar{s}}{\sqrt{\frac{\sum_{i=1}^{n}(s_i - \bar{s})^2}{n-1}}}$$

2.6 Baseline Correction

Baseline correction is used to change spectral offset by subtracting either a spectral value at a specified wavelength, or a constant value entered manually.

2.7 Detrend

Detrend is a method that can be used to remove baseline offset, slope, or curvature from a spectrum. This is accomplished by calculating a baseline function as the least squares fit of a polynomial to the sample spectrum, and then subtracting that function from the spectrum.

The polynomial function models the effects in a cumulative fashion as its order increases from 0^{th} to 1^{st} and 2^{nd} degrees:

Order of the Polynomial Function	Baseline Effect(s) Removed
0	offset
1	offset and slope
2	offset, slope, and parabolic curvature

2.8 Savitzky-Golay

This well known method of smoothing and derivative calculation relies on the least squares fit of a polynomial to a spectral segment. Though both the Savitzky-Golay (S-G) and Detrend methods are based on the least squares fit of polynomial functions, they differ in both scope and effect:

Method	What is modeled	The Result
Detrend	Broad features across the whole spectrum, i.e. offset, slope and curvature (a single polynomial function is fit to the entire spectrum)	The modeled function is subtracted from the spectrum
Savitzky-Golay	Fine spectral features defined by fitting a polynomial function to a limited number of data points centered on a single data point (the data points in the segment are modeled by a polynomial function)	The modeled value at the central wavelength replaces the original spectral value at that point

In the S-G method, a smoothed spectrum or a derivative spectrum of any order can be calculated using coefficients of the polynomials fitted to a spectrum.

Three parameters must be defined to apply the S-G pretreatment:

S-G Parameter	Purpose
Number of points	Defines the size of the convolution window (the size of the spectral segment to which the polynomial is fit)
The order of the convolution polynomial	Determines the complexity of the polynomial function presumed to be contained in the convolution window
Method outcome	Smoothed spectrum; or 1^{st}, 2^{nd} and 3^{rd} derivatives

2.9 Multiplicative Scatter Correction

All math pre-treatments discussed so far are based on and applied to individual spectra: they operate on data points in a given spectrum, and yield results determined by the unique characteristics of that spectrum. By contrast, Multiplicative Scatter Correction (MSC) is set-dependent: it is a scatter correction method based on a related set of spectra.

In MSC, the mean spectrum is calculated from all spectra in a defined data set. Then a least squares

linear regression is performed on absorbance values of the sample spectrum versus those at corresponding wavelengths in the mean spectrum. This operation yields a linear equation with a defined intercept and slope. Next, the value of the intercept is subtracted from every data point in the spectrum. Finally, each absorbance value in the resulting spectrum is divided by the value for slope. Using the mean spectrum, this same set of operations is performed on every spectrum in the data set.

Pretreatments like derivatives, SNV, Detrend, and Savitsky-Golay are applied to individual spectra without any preconception of what the resulting spectrum should look like. By contrast, the MSC method calculates a mean spectrum under the assumption that spectra in the data set are distributed normally. Thus, the mean spectrum is the most probable spectrum, i.e. it is the highest probability representation of all spectra in the data set.

MSC effectively coerces individual spectra to behave like the mean spectrum as much as possible. Success of the method strongly depends on the calculated mean spectrum closely resembling the true mean spectrum, which depends in turn on a large sample set. Quantitative method development usually requires such data sets to completely span the full range of chemical and spectral variation encountered when the method is actually implemented.

Sample sets for qualitative models development also may be large – indeed, they may include many times more samples than quantitative models – because they often include dozens of hundreds of products. However, a given product may contain only a few samples. Consequently, MSC may not be as suitable for qualitative analysis as it is for those quantitative applications where elimination of pathlength and slope variation in a single pretreatment is important.

2.10 Thickness Correction/ normalization (pathlength correction)

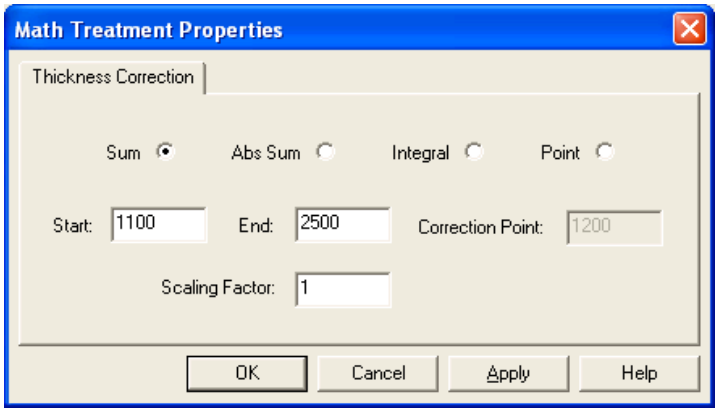
Vision makes Thickness Correction/normalization available to the user as a math pretreatment of spectra. Thickness Correction, or pathlength correction, is useful for spectral measurements of samples where the pathlength is not guaranteed to be constant. It can also be used to address changes in density due to temperature variations. Mathematically, Thickness Correction can be applied to transmission and reflectance measurements.

It is applicable to quantitative measurements of liquids where a single absorption band can be identified for the constituent of interest such as MLR regression and an isolated band in every spectrum that is attributable to a constituent that does not vary in concentration for all samples. That isolated band can then be assumed to be entirely related to the increase or decrease in the samle pathlength. By normalizing the entire spectrum to the intensity of that band, pathlength variation is effectively removed. When the only bands in the region being used for quantitative analysis are from a single constituent of interest, then thickness correction/normalization is not beneficial as it will result in a constant intensity. However, when there are two or more constituents of interest with differing absorbtivities, Thickness Correction/Normalization is very powerful.

If applied in a PLS regression, where a wavelength region is used to calculate the integrated area for normalizing the region, the range of the constituent of interest concentrations must be relatively small. This is because large variations in concentrations would cause the integrated areas to vary mostly by concentration, not pathlength differences. This would introduce non-linearities in the constituent to spectra correlations and reduce the prediction accuracy of the model.

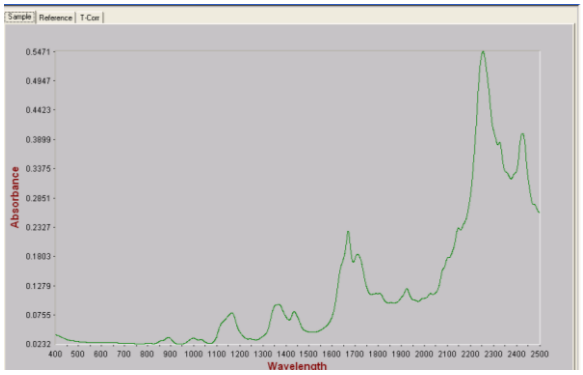
In Vision - Thickness Correction can be selected before or after other treatments when choosing math pretreatments.

Sum

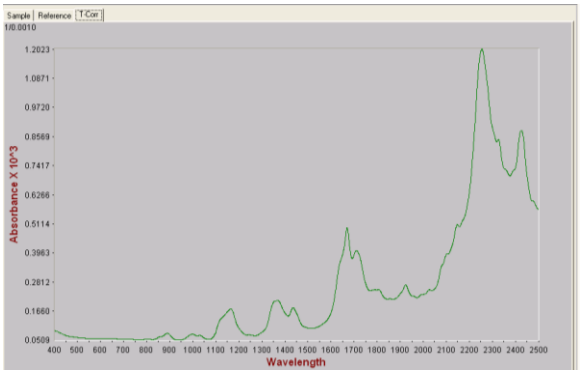


The absorbance values over the entire region of the spectrum would be divided by the Sum of the absorbance values over the wavelength region 1100 – 2500nm.

Before

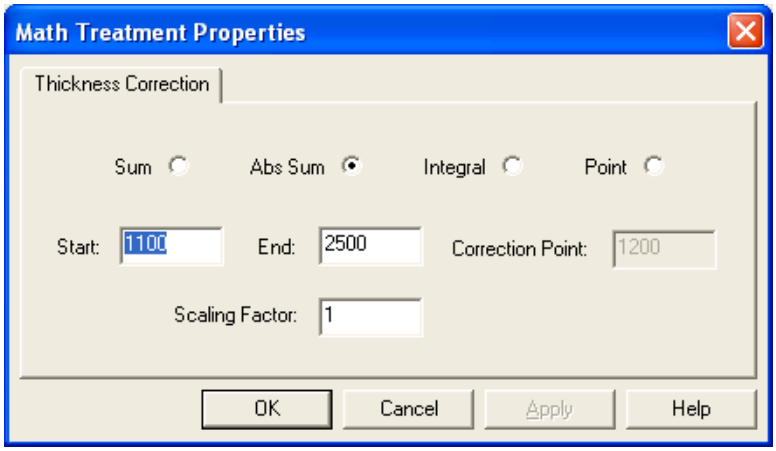


After



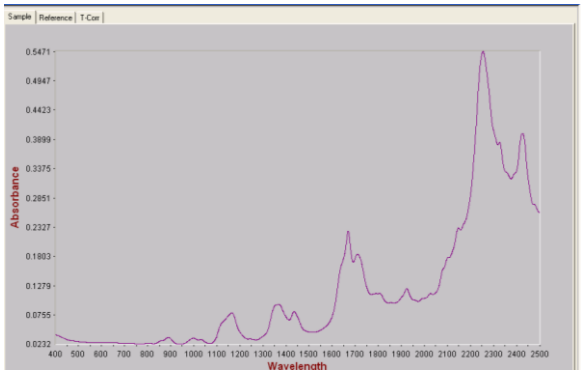
Notice the difference in the y axis scale.

Absolute Sum

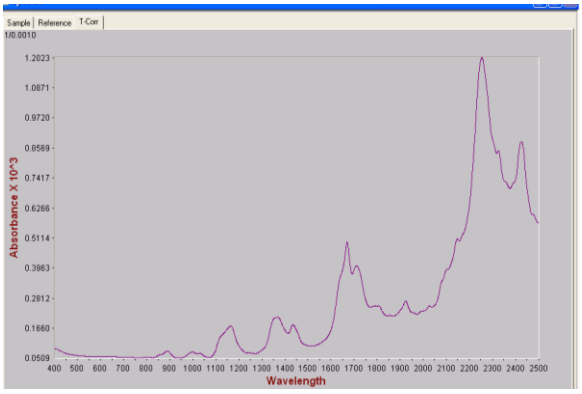


The absorbance values over the entire region of the spectrum would be divided by the Absolute Sum of absorbance values over the wavelength region 1100-2500nm.

Before

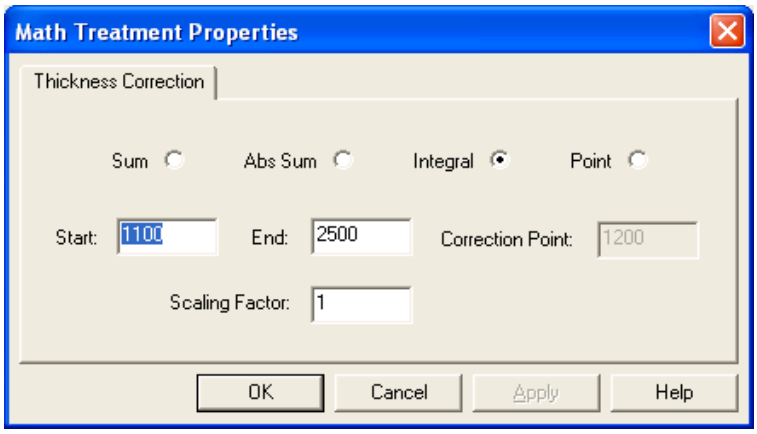


After



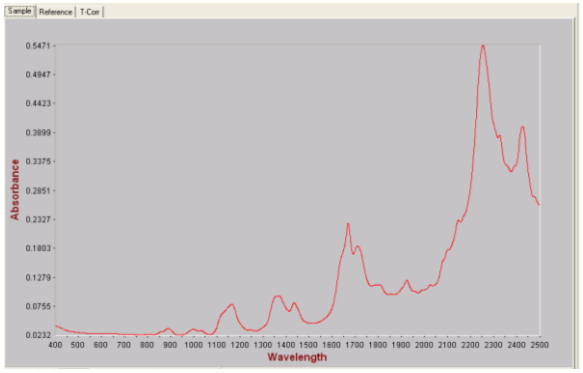
Notice the difference in the y axis scale.

Integral

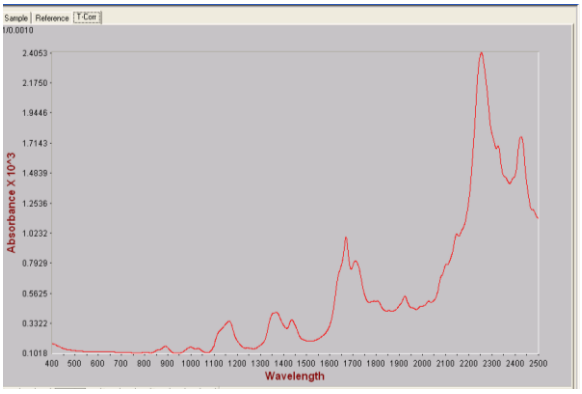


The absorbance values over the entire region of the spectrum would be divided by the Integral (area) of the spectrum over the wavelength region 1100-2500nm.

Before

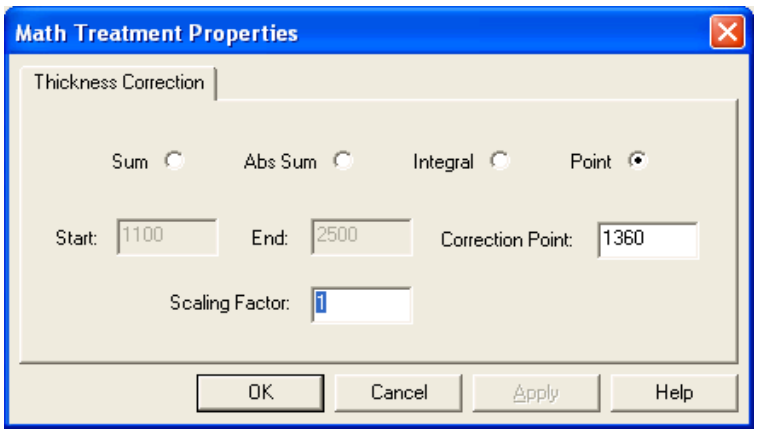


After



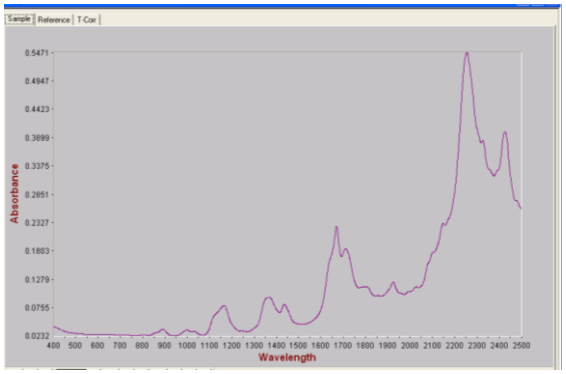
Notice the difference in the y axis scale.

Point

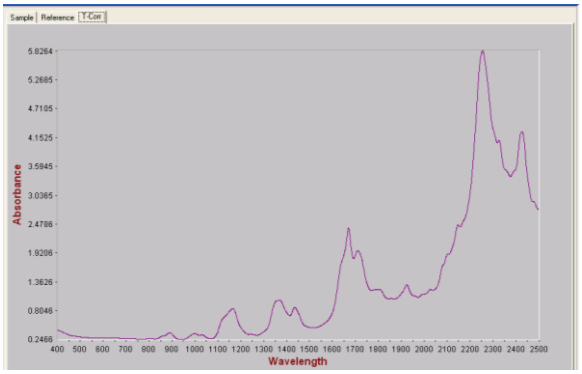


The absorbance values over the entire wavelength region of the spectrum would be divided by the absorbance value at Point (wavelength) 1360 nm.

Before

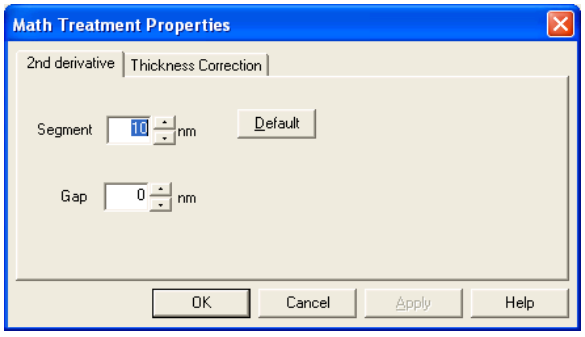


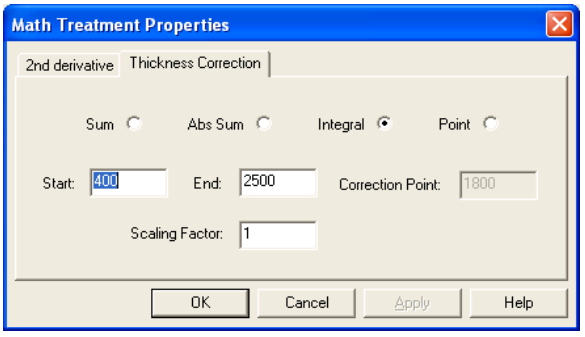
After



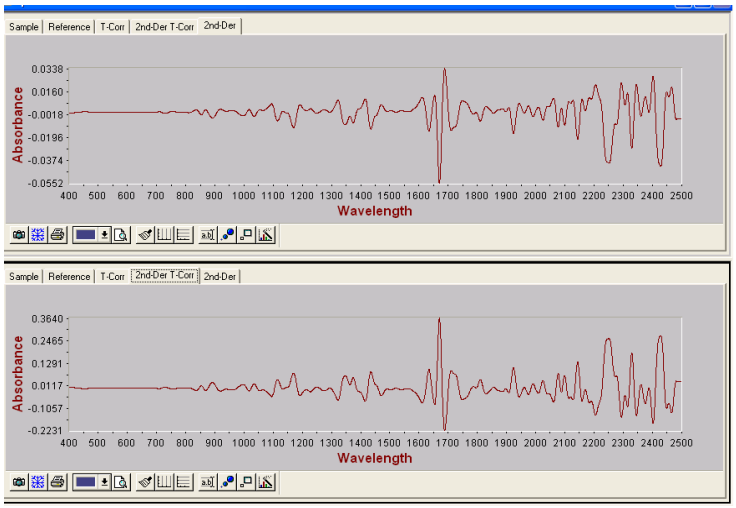
Notice the difference in the y axis scaling.

Visual differences are very apparent when thickness correction is applied with a derivative math treatment.





The top graph is the sample with only a second derivative math treatment (2/10/0) while the lower graph uses the same 2nd derivative treatment plus the thickness correction using the integral over the 400-2500nm range.



3 Qualitative Library Development

3.1 Principal Component Analysis

The near IR spectrum comprises intensity measurements at hundreds of wavelengths. Because there is always some correlation among absorbance values within a spectrum, the information contained in spectra for a set of related materials is all the more redundant. Thus, while there are many wavelengths, there may be relatively little unique spectral information.

Principal Component Analysis (PCA) is a method capable of describing the unique variances in a set of spectra using linear combinations of wavelengths (Principal Components, or PCs). These PCs are uncorrelated and, because information in the data set is redundant, relatively few are required to account for the significant information in the spectral data set.

Mathematically, the Principal Component Analysis is performed by calculating a set of eigenvectors that diagonalize the covariance matrix C of the training set of spectra:

$$C = EDE^{-1}$$

Where ***C*** is the covariance matrix, ***E*** is the Eigen vectors matrix, and ***D*** is a square diagonal matrix with Eigen values. The Eigen vectors have the same length as spectra in the training set, and are orthonormal. From this property it follows that:

$$E^{-1} = E^{T}$$

There are numerous algorithms that can be used for Eigen value decomposition of a matrix. Vision software uses a well established and numerically stable algorithm called **Singular Value Decomposition**. The algorithm operates on a mean-centered training set of spectra and returns a set of eigenvectors together with associated Eigen values. Eigen vectors are arranged in the decreasing order of Eigen values.

If all possible Eigen vectors were included in a PC model, it would account for 100% of the variance in the spectral training set. This would happen if ***N-1*** PCs were used (***N*** is the number of spectra in the training set, and one degree of freedom is lost due to mean-centering). Normally, only a few PCs (the **primary PCs**) will account for the majority of variance in the training set. Remaining principal components (called **secondary PCs**) are usually attributed to noise.

A quantity called cumulative variance shows what is a percent of total variance described by a given number (***m*** in this case) of PCs:

$$V_c = \frac{\sum_{i=1}^{m} \lambda_i^2}{\sum_{i=1}^{N-1} \lambda_i^2}$$

Where λi is the Eigen value that corresponds to the ith Eigen vector. The default value of cumulative variance is 95 %.

Principal components can be interpreted as the axes of a new, orthogonal system of coordinates. Multiplication of a spectrum by an eigenvector yields a number, called a score:

$$s_i = AE_i$$

Multiplication of a spectrum by a set of eigenvectors yields a set of scores, which can be interpreted

as coordinates of the spectrum in the principal component space:

$$s = AE$$

It is also possible to reconstruct a spectrum from its principal component scores, knowing the transformation eigenvectors:

$$A_r = sE^T$$

If the transformed spectrum belonged to the training set, and the number of principal components in the model was such that 100 % of the variance was accounted for, then the reconstructed spectrum is identical to the transformed spectrum. Normally, since only the primary principal components are used for reconstruction, the spectra are slightly different. The difference is called the residual spectrum:

$$R = A - A_r$$

The variance of the residual spectrum:

$$V_r = R^T R$$

Can be used as an indicator whether the spectrum belongs to the same distribution as the training set spectra in so called Residual Variance method.

3.2 Distance Metrics

3.2.1 Euclidean Distance

Euclidean distance between two objects (e.g., spectra) ***x*** and ***y*** with ***n*** coordinates (wavelengths or principal component scores) is calculated according to the following formula:

$$D_E = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

3.2.2 Mahalanobis Distance

The Mahalanobis distance, on the other hand, is the distance between a spectrum ***A*** and the center of the distribution of a set of spectra represented by a covariance matrix ***C*** as:

$$D_M^2 = (A - \mu)^T C^{-1} (A - \mu)$$

Where μ denotes the mean spectrum of the distribution. The spectrum ***A*** may or may not belong to the training set. Calculating the Mahalanobis distance for a spectrum in this way has disadvantages. An alternative, simpler approach calculates Mahalanobis distance from primary principal component scores after secondary PCs have been rejected from the model:

$$D_{M,i}^2 = (n-1)^2 \sum_{j=1}^{k} \left[s_{ij}^2 \sum_{i=1}^{n} s_{ij}^2 \right]$$

Where the index ***j*** runs over all ***k*** principal components used for the model, and the index ***i*** runs over all ***n*** samples in the training set.

3.2.3 Maximum Distance

To calculate maximum distance (distance in this context is in terms of spectral value – absorbance, or

the spectral value in a derivative spectrum) of a spectrum ***y*** to a mean product spectrum ***x***, Vision first defines the inflated standard deviation spectrum from the product set:

$$s_i^d = \left(1 + \frac{1}{\sqrt{2(n-1)}}\right)\left(\frac{\sum_j (x_j - \overline{x_j})^2}{n-1}\right)^{\frac{1}{2}}$$

Where the index i runs over all wavelengths, and index ***j*** runs over all spectra in the set.

Consequently, the maximum distance is calculated as:

$$D_x = \max\left[abs\left(\frac{y_i - \overline{x_i}}{s_i^d}\right)\right]_{over\ all\ i}$$

Maximum distance can be interpreted as the maximum deviation from the product mean spectrum expressed in units of inflated standard deviation.

3.2.4 Correlation

Correlation formally is not a distance, but a measure of spectral similarity. Correlation between two spectra ***x*** and ***y*** is calculated as:

$$D_c = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

Strictly speaking, correlation is a dot product of two vectors representing spectra. Geometrically, it is a cosine of an angle between those vectors. Correlation is scale invariant, i.e. its value does not change if one or both spectra are multiplied by a constant.

4 Sample Selection Methods

4.1 Mahalanobis Distance in Principal Component Space

4.1.1 Outlier Detection

Principal Component Analysis performed on spectra for a given product yields a set of eigenvectors with corresponding eigenvalues. From the cumulative variance threshold defined for the model, the number of PCs in the model is determined. Multiplication of spectra and eigenvectors yields scores (spectral coordinates in PC space).

Calculation of the Mahalanobis distance is done on PC scores of all product spectra. Assuming that the spectra in the training set are distributed normally, Mahalanobis distances are distributed according to chi-square function. Chi-square is a function with well-known properties. From the chi-square function one can calculate probability that a given sample belongs to the distribution represented by the training set.

The method's threshold defines boundaries of the model ellipsoid. During analysis, samples outside the ellipsoid will fail identification or qualification. The threshold can be of two types: **probability** or **match value**.

Threshold expressed as **probability** is the recommended type. Vision has built-in chi-square distribution function. A sample's Mahalanobis distance and the number of degrees of freedom of the training set is passed to this function, which returns a probability that the sample does not belong to the distribution represented by the training set of spectra.

The chi-square distribution (and consequently Mahalanobis distance value for samples in the training set) depends strongly on the number of samples in the training set. For example, if a training set contains hundreds of spectra, the Mahalanobis distance value is expected to exceed one hundred for even good samples. For this reason, scaling is usually performed by dividing a mean value into the distance value.

However, Vision uses a different scaling factor, the number of degrees of freedom. The Mahalanobis distance scaled in this way is used when the **Match Value** is the type of outlier threshold chosen. The default value, 0.6, is not statistically meaningful and has been established experimentally. Samples with scaled Mahalanobis distance above this value will be tagged as outliers.

4.1.2 Redundant Samples Selection

Redundant samples are detected based on Euclidean distances in PC space (calculated on PC scores). After removal of outlier samples, remaining samples undergo redundant sample selection.

If the distance threshold method is used to select redundant samples, Vision randomly picks a spectrum and calculates distances from this spectrum to all other spectra. This spectrum is placed in the training (or calibration) set, and all spectra with distances smaller than the threshold are placed in the acceptance (validation) set. The process continues until all spectra are distributed between appropriate sets.

Because the calculated distances are not scaled, threshold values depend on the product spectra. Therefore, to optimize sample selection for a given product, several runs may be required. For this reason, **By Number of Samples** is the preferred option for sample selection. In this case Vision

estimates the what threshold is required to obtain the desired distribution of samples between training and acceptance sets.

Before the sample selection is performed, the data set may include outliers. Because the presence of outliers distorts the Principal Component model, optimally the outliers should be removed from the set as soon as they are detected, and the model recalculated before redundant samples are found. This can be done by selection the Reset Product Mean option during Sample Selection.

If this is the case, Vision will display on all plots the results of the second method, without outliers. However, if the outliers are found, they are saved in the Rejection Set.

If you want to see the outliers on plots, do not select the Reset Product Mean option. However, be aware that redundant samples search is done within the model that may include outliers.

4.2 Maximum Distance in Wavelength Space

4.2.1 Outlier Detection

In this method outliers are detected by calculating maximum distance of all product spectra from the product mean spectrum. The standard deviation envelope is defined using all product spectra.

4.2.2 Redundant Sample Selection

Redundant samples are detected based on Euclidean distances in wavelength space (calculated on spectra). After removal of outlier samples, remaining samples undergo redundant sample selection.

If the distance threshold method is used to select redundant samples, Vision randomly picks a spectrum and calculates distances from this spectrum to all other spectra. This spectrum is placed in the training (or calibration) set, and all spectra with distances smaller than the threshold are placed in the acceptance (validation) set. The process continues until all spectra are distributed between appropriate sets.

Because the calculated distances are not scaled, threshold values depend on the product spectra. Therefore, to optimize sample selection for a given product, several runs may be required. For this reason, By Number of Samples is the preferred option for sample selection. In this case Vision estimates the what threshold is required to obtain the desired distribution of samples between training and acceptance sets.

Before the sample selection is performed, the data set may include outliers. Because the presence of outliers distorts the Principal Component model, optimally the outliers should be removed from the set as soon as they are detected, and the model recalculated before redundant samples are found. This can be done by selection the Reset Product Mean option during Sample Selection.

If this is the case, Vision will display on all plots the results of the second method, without outliers. However, if the outliers are found, they are saved in the Rejection Set.

If you want to see the outliers on plots, do not select the Reset Product Mean option. However, be aware that redundant samples search is done within the model that may include outliers.

4.3 Random Selection

When this method of sample selection is chosen, Vision randomly splits all the product spectra into training and acceptance sets. No rejection (outlier) set is created when this method is applied.

4.4 Sample Selection Based on Lab Data (Quantitative)

In this method of sample selection Vision displays a histogram of lab data distribution. The distribution can be modified in attempt to approach a boxcar distribution.

5 Identification and Qualification Methods

5.1 Wavelength Correlation

5.1.1 Model Development

The first step in development of a wavelength correlation model is the calculation of a mean spectrum for every product in the training set. Every sample spectrum for a product is used in the calculation of the mean spectrum for that product.

5.1.2 Analysis of an Unknown

The correlation between an unknown spectrum and every mean product spectrum is calculated. The unknown is identified as that product for which the correlation value is above the threshold value. The default thresholds are 0.84 for identification and 0.9 for qualification.

5.2 Wavelength Maximum Distance

5.2.1 Model Development

The first step in development of a wavelength correlation model is the calculation of a mean spectrum for every product in the training set. Every sample spectrum for a product is used in the calculation of the mean spectrum for that product.

5.2.2 Analysis of an Unknown

The maximum distance between an unknown spectrum and every mean product spectrum is calculated. The unknown is identified as that product for which the correlation value is below the threshold value. Default thresholds are 4 for identification and 3 for qualification.

5.3 Mahalanobis Distance in Principal Component Space

Library identification based on Mahalanobis distance calculates a local Principal Component model for each product in the library. Then a qualification method is developed separately for each product.

5.3.1 Model Development

Principal Component Analysis performed on the training set of spectra of a given product yields a set of eigenvectors with corresponding eigenvalues. From the cumulative variance threshold defined for the model, the number of primary PCs in the model is determined. Multiplication of spectra and eigenvectors yields scores (spectral coordinates in the Principal Component space).

Mahalanobis distances are distributed according to chi-square function. (Chi-square is a statistical function with well-known properties.) From the chi-square function one can calculate probability that a given sample belongs to the distribution represented by the training set.

The Mahalanobis distance method offers a choice of two types of thresholds : probability or match value. Threshold expressed as probability is the recommended type. Vision has built-in probability function based on the chi-square distribution. Samples' Mahalanobis distances and the number of degrees of freedom of the training set are passed to this function, which returns a probability that the sample does not belong to the distribution represented by the training set of spectra ($1-\alpha$).

Another type of threshold is match value. Mahalanobis distance calculated directly from the formula depends strongly on the number of samples in the training set. Therefore if the match value type of threshold is to be used, the distance has to be scaled. Vision divides Mahalanobis distance by the number of degrees of freedom. Default threshold values have been determined experimentally and do not have statistical interpretation.

5.3.2 Analysis of an Unknown

The Mahalanobis distance between an unknown spectrum and a product mean spectrum is calculated using primary eigenvectors. The distance is scaled or the probability calculated (depending on the choice of threshold). The unknown passes analysis if the calculated distance is below the threshold value.

Note: If the probability threshold is used, Vision calculates the probability that a sample is not a member of the distribution described by the training set of spectra. A low value for this quantity indicates a high probability that the sample spectrum belongs to the training set.

5.4 Residual Variance in Principal Component Space

Library identification method based on residual variance calculates a local Principal Components model for each product in the library. Qualification method is developed for each product separately.

5.4.1 Model Development

Principal Component Analysis performed on spectra in the training set of a given product yields a set of eigenvectors with corresponding eigenvalues. From the cumulative variance threshold defined for the model, the number of primary PCs in the model is determined.

Residual variance is distributed according to F function. (F distribution is a ratio of two chi-square distributions). From the F function one can calculate probability that a given sample belongs to the distribution represented by the training set.

Residual variance method offers a choice of two types of thresholds: probability or match value. Threshold expressed as probability is the recommended type.

Vision has built-in probability function based on the F distribution. Samples' residual variance and the number of degrees of freedom is passed to this function, which returns probability that the sample does not belong to the distribution represented by the training set of spectra ($1-\alpha$).

Another type of threshold is match value. The calculation of residual variance from the formula is not intuitive. Therefore if the match value type of threshold is to be used, the variance has to be scaled. Vision scales residual variance by the number of degrees of freedom. The default threshold values have been determined experimentally and do not have statistical interpretation.

5.4.2 Analysis of an Unknown

Using a products' principal component model, the residual variance of the unknown spectrum is calculated using primary eigenvectors. The variance is scaled or the probability calculated (depending on the choice of threshold). The unknown passes analysis if the calculated residual is below the threshold value.

Note: If the probability threshold type is used, Vision calculates the probability that a sample is not a member of the distribution described by the training set. A low value for this quantity indicates a high probability that the sample spectrum belongs to the training set.

6 Library Clustering

6.1 General Description

Library clustering is an algorithm designed to generate a simplified representation of a large, diverse library. At the first level of clustering, a library is separated into clusters that contain similar spectra. At the second level, each of these clusters is in turn subdivided into sub-clusters. This process can continue until clusters can no longer be subdivided by application of a set of clustering parameters, or until the maximum number of clustering levels (32) has been reached. The lowest, non-divisible level clusters are called leaf clusters.

6.2 The Minimal Spanning Tree Algorithm

To describe the mechanism of clustering, it is necessary to understand the basic algorithm used in clustering, the minimal spanning tree algorithm.

Imagine a library with n products, each represented by a mean product spectrum. The number of all possible distances between products is of the order of n^2. The purpose of the minimal spanning tree algorithm is to reduce the number of distances to the most significant products.

1. The algorithm starts with all products in a bin designated A.
2. A product is randomly chosen as a current product and moved to a bin designated B.
3. The distances from the current product to all remaining products in Bin A are calculated.
4. The shortest distance from current product to all remaining products in Bin A is determined.
5. The product for which the shortest distance was found is made the current product and moved to Bin B. The distance value and the names of products to which it connects are saved. If Bin A is empty, the algorithm stops here.
6. The distances from the current product to all remaining products in the Bin A are calculated. If the shortest distance is smaller than the distance saved in (5.), go to (4.).
7. If the shortest distance is longer than the distance saved in (5.), move back to the last current product, make it current product and go to (4.).

The result of the algorithm is a minimal spanning tree, which is a representation of the library with the following properties:

- The number of distances between library products is reduced from n(n-1)/2 to n-1.
- The sum of those distances is the smallest for all possible trees.
- The final result of the algorithm does not depend on the choice of the starting point.

6.3 Clustering Algorithm

Three parameters must be defined to create a clustering method:

1. PC cumulative variance
2. Variance radius scale
3. The maximum leaf cluster size

Once these parameters have been defined, the algorithm proceeds as follows:

1. A principal component model is calculated for the library using that number of primary PCs required to satisfy the cumulative variance value (parameter #1).
2. PC scores are calculated for all mean product spectra.
3. The standard deviation spectrum (in wavelength space) is calculated for each product.
4. The Euclidean norm is calculated for each product's standard deviation spectrum.
5. The minimal spanning tree is determined using Euclidean distances on principal component scores of product mean spectra.
6. Spheres are drawn around each product, with radii equal to the Euclidean norm of the product standard deviation spectrum multiplied by the variance radius scale (parameter #2).
7. If there is overlap between spheres for any pair of products on the tree, the products end up in one cluster.

The algorithm described above may result in separation of a library into a number of clusters. Each of those clusters in turn will undergo the same procedure in the next level of clustering (based on a PC model local to the cluster) if the number of products in this cluster is equal or larger than the maximum leaf cluster size (parameter #3).

The last step in the clustering procedure is to define the cluster boundaries. All leaf clusters are enclosed by rectangular boxes. Each box has a size sufficient to enclose all the product spheres in a cluster. (Radii of the spheres are equal to Euclidean norm of standard deviation multiplied by the variance radius scale parameter.) The coordinates of the center of each cluster box and dimensions are saved with the method.

6.4 Analysis of an Unknown

PC scores for the unknown spectrum are calculated from the first clustering level PC model. The location of the unknown spectrum in the PC space is determined. The unknown spectrum may be identified as one of the products if 1) its location locates within boundaries of any of the first level clusters defined during clustering method development, and 2) that cluster is not further subdivided into clusters.

If the location does not fit any of the clusters, the unknown fails clustering (is declared not present in the library). If the cluster into which the unknown falls is further subdivided, the second level PC model is applied to the unknown spectrum to determine if it falls into one of subclusters. This process continues until a leaf cluster level is reached.

If the unknown spectrum locates within the boundaries of one of the leaf clusters, an identification method local to this cluster is applied and the leaf cluster is searched to identify the unknown sample.