

## Chemometrics

**Chemometrics:** Mathematical, statistical, graphical or symbolic methods to improve the understanding of chemical information. – or – The science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods

Chemometrics often involves using linear algebra methods to make qualitative or quantitative measurements of chemical data.

Chemometrics solves an individual problem but do not address ALL possible problems.

Some methods have the advantage of being simple to understand but may not be very robust for all possible samples. Others are very complex to understand and implement, but give solutions that are very stable and can handle a large variety of "unknowns."

### N-dimensional space

Extension of regular three-dimensional space to n-dimensions, reflecting more complex multivariable situations.

Practical problems are often solved using more than three dimensions with cluster analysis and pattern recognition.

### Projection and Mapping

- Reduction of dimensions is a primary goal of multivariate analysis.

*Two approaches:*

1. Find an important subset of the original variables.
2. Synthesize new variables from the original variables.

The creation of new variables can be approached in one of two ways: **projection** and **mapping**.

### Projection

Linear combinations of the original variables are created which can define a new, smaller set of variables, while retaining as much information as possible. Principal components analysis (PCA) is an example.

### Mapping

Mapping transformations are non-linear.

They preserve special properties of the data (e.g. interpoint distances), while performing data reduction.

The results from mapping can be difficult to interpret.

As with statistics, the key to understanding chemometrics is not necessarily understanding the mathematics of all of the different methods; it is knowing which model to use for a given analytical problem and properly applying it.

Methods:

[Beer-Lambert Law](#)  
[Least Squares Regression](#)  
[Classical Least Squares \(K-Matrix\)](#)  
[Inverse Least Squares \(P-Matrix\)](#)  
[Partial Least Squares \(PLS\)](#)  
[Discriminant Analysis](#)  
[Preprocessing Techniques](#)  
[Principal Components Regression \(PCR\)](#)  
[Qualitative Spectroscopy Methods](#)  
[Quantitative Spectroscopy Methods](#)

## The Beer Lambert Law

One of the keys to quantitative analysis in any scientific field is the assumption that the amounts (concentrations) of the constituents of interest in the samples are somehow related to the data from a measurement technique used to analyze them. The ultimate goal is to create a calibration equation (or series of equations) which, when applied to data of "unknown" samples measured in the same manner, will accurately predict the quantities of the constituents of interest.

In order to calculate these equations, a set of "standard" samples are made which reflect the composition of the "unknowns" as closely as possible. These standards are designed to span the expected range of concentrations and compositions in the unknowns and are measured under the same conditions as the unknowns. The standards are then measured by an instrument. Together, this collection of known data (the composition of each standard) and the measured data from the instrument form what is known as a training set or calibration set. The calibration equations that describe the relationships between these two sets of information are calculated from this data. The exact equation or set of equations that make up the calibration is also known as a model. Thus, this process is often called, "solving the calibration model."

Once the model equations have been selected and solved, they can be used to calculate the same quantities or properties in "unknown" samples. However, in order for these sample(s) to be predicted accurately, they must be measured under exactly the same conditions on the same instrument as the calibration set.

For many current applications, a spectrometer is increasingly becoming the measurement device of choice. Unlike other methods which give "single point" measurements for each calibration and unknown sample (i.e., pH, or single element Atomic Absorption), the spectrum of a sample contains many data points. Every response value in a spectrum has some relation to the properties or constituent(s) that make up the measured sample. Using a spectrum of a sample that has many data points has some distinct advantages over single point measurement techniques. One of the most important factors is that there are many more measurements per sample (spectral data points) to use in generating the calibration equations. As anyone who has performed quantitative analysis knows, the more measurements per sample, the more accurate the results. The problem for the analyst is to discover what those relationships are, and use a calibration model that reflects them accurately.

One advantage of using spectroscopy as a measurement technique is that the Beer-Lambert Law (also known as Beer's Law) defines a simple linear relationship between the spectrum and the composition of

a sample. This law, which should be familiar to all spectroscopists, forms the basis of nearly all other chemometric methods for spectroscopic data. Simply stated, the law claims that when a sample is placed in the beam of a spectrometer, there is a direct and linear relationship between the amount (concentration) of its constituent(s) and the amount of energy it absorbs. In mathematical terms:

$$A_{\lambda} = \epsilon_{\lambda} bC$$

Where  $A_{\lambda}$  is the sample's Absorbance value at specific wavelength (or frequency)  $\lambda$ ,  $\epsilon_{\lambda}$  is the absorptivity coefficient of the material (constituent) at that wavelength,  $b$  is the pathlength through the sample and  $C$  is the concentration. The absorptivity coefficient for every material is different, but for a given compound at a selected wavelength, this value is a constant. The only problem that remains is to discover the "real" value of that constant.

**Least Squares Regression** Quantitation by spectral analysis can be done using a variety of mathematical techniques. This technique addresses the simplest of these methods which is based on the direct measurement of band height or area from one or more spectra. At times the analyst is only interested in the raw numeric values of this measurement which is then compared to previous measured values as in a quality assurance procedure. The raw value may also be scaled by a single concentration factor which relates the raw value directly to the concentration values of the analysis. A more accurate and advanced technique of spectral quantitative analysis is to create a calibration equation or series of equations which, when applied to spectra of "unknown" mixture samples, will accurately predict the quantities of the components of interest. In order to calculate these equations, a set of "standard" mixtures are made which reflect the composition of the "unknowns" as closely as possible. These standards are also designed to span the expected range of concentrations and compositions in the unknowns and are measured under the same conditions (sampling method, pathlength, instrument, etc.) as the unknowns. The spectra of the standards (in [absorbance](#) or other concentration-proportional units) are then measured by a spectroscopic instrument and saved in digital format. This set of spectra and the known quantities of the components in each individual sample form what is known as a training set or calibration set from which the calibration equations are built. The unknown sample(s) are then measured in the same manner on the same instrument and the equations are used to "predict" the concentration of the calibrated components in each unknown.

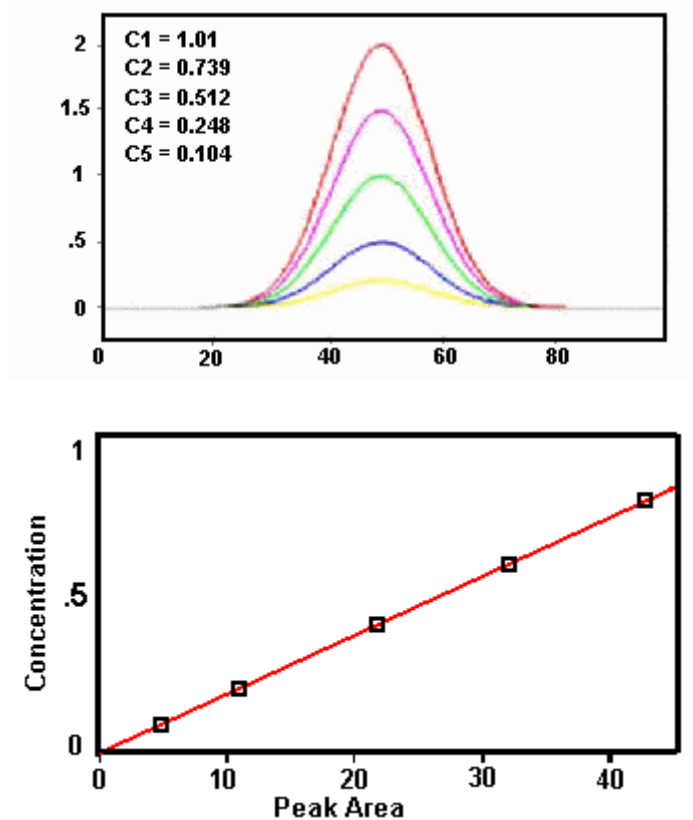
One of the keys to quantitative analysis is the assumption that the concentrations of the components of interest are somehow related to the data from the measurement technique used for analyzing the samples. This relationship must be able to be accurately described by a calculated equation in order to be useful for predicting the compositions of any "unknown" samples measured. In some cases, the components of interest may have well-resolved bands. In these cases, either the peak height or the peak area can be related to concentration by a single simple equation. These equations can take the form of a straight line or even a quadratic curve. In this case, concentrations of the components are calculated from equations such as:

$$C_a = B_{1,a} (Area_a) + B_{0,a}$$

$$C_a = B_{2,a} (Height_a)^2 + B_{1,a} (Height_a) + B_{0,a}$$

where  $C_a$  is the concentration of component  $A$ , and the  $B_s$  are the calibration coefficients. By measuring the height or area of a component band in the spectrum, it is possible to compute the concentration using the equation and these coefficients.

However, the coefficients are not necessarily known ahead of time and must be calculated. This is accomplished by first measuring the spectra of some samples of known concentration. As you can see from the equations, there are either two or three unknown coefficients. This means at least two or three known samples must be measured in order to solve the equation. However, more are usually measured to improve the accuracy of the calibration coefficients. In fact, sometimes multiple runs of the same concentration are used to get an averaging effect.



The areas of spectral component band and the component concentrations (top) were used to compute the coefficients of the calibration equation by Least Squares Regression (bottom).

The [peak areas](#) or [heights](#) of the component band and the known concentrations of all the calibration spectra can then be used to calculate the coefficients. The best way to find the calibration coefficients from this set of data is by a Least Squares Regression. This is a mathematical technique that calculates the coefficients of a given equation such that the differences between the known responses (peak areas or heights) and the predicted responses are minimized. (The predicted measurements are those calculated by reading the measurements off of the line at the known concentrations.) If there is more than one component in the samples, a separate band must be used for each component of interest. This also means one equation is necessary for each component. Once the equations are calculated for each component, the concentrations of these components in "unknown" samples can be calculated by

substituting the peak areas or peak heights into the proper equation for that component band and solving for concentration.

While this method is conceptually easy to understand and the calculations are straightforward, it will not produce accurate results for mixtures with overlapping bands. Since Least Squares Regression assumes that the absorbance measurement of peak height or total peak area is the result of only one component, the predictions will have large errors if there are interfering components that have spectral bands that overlap those of the components of interest. In these cases, more sophisticated mathematical techniques are necessary such as [Inverse Least Squares \(ILS\)](#), [Partial Least Squares\(PLS\)](#), [Principal Component Analysis \(PCA\) Methods](#), or [Principal Component Regression\(PCR\)](#).

## Classical Least Squares (CLS)

This spectroscopic quantitation method, also known as **K-Matrix**, is founded in using the [Beer-Lambert Law](#) to extend the calculation of the absorptivity coefficients across a much larger portion of the spectrum than the much simpler [Least Squares Regression](#) method. Referring back to the description of Beer's Law, notice that it defines a relationship between 4 different variables; the spectral response ( $A_\lambda$ ), the constituent absorptivity constant ( $\epsilon_\lambda$ ), the pathlength of light ( $b$ ) and the constituent concentration ( $C$ ). The goal of calibrating a spectroscopic quantitative method is solving for the absorptivity constants. However, if the pathlength of the samples is also kept constant (as it is for most quantitative experiments), Beer's Law can be rewritten as:

$$A_\lambda = K_\lambda C$$

where the absorptivity coefficient and pathlength are combined into a single constant,  $K$ . This equation can be easily solved by measuring the absorbance of a single sample of known concentration and using these values to solve for  $K_\lambda$ . Predicting the concentration of an unknown sample is as simple as measuring the Absorbance at the same wavelength and then applying the following:

$$C = \frac{A_\lambda}{K_\lambda}$$

However, basing an entire calibration on a single sample is generally not a good idea. Due to limitations of noise, instrument error, sample handling error, and many other possible variations, it is best to measure the absorbances of a series of different concentrations and calculate the slope of the best fit line through all the data points. Just as in the case of [Least Squares Regression](#), this can be solved with a simple regression line of absorbance versus concentration.

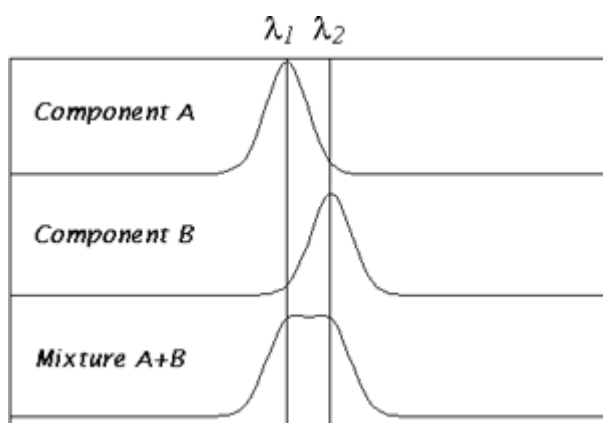
However, the problem becomes more complex if the sample contains two constituents. In any algebraic solution, it is necessary to have as many equations as unknowns. In this case, it is necessary to set up two equations:

$$\begin{aligned} A_{\lambda 1} &= K_{\lambda 1} C_a \\ A_{\lambda 2} &= K_{\lambda 2} C_b \end{aligned}$$

where  $A_{\lambda_1}$  and  $A_{\lambda_2}$  are the absorbances at two different wavelengths,  $C_a$  and  $C_b$  are the concentrations of the two constituents ("A" and "B") in the mixtures, and  $K_{a,\lambda_1}$  and  $K_{b,\lambda_1}$  are the absorptivity constants for the two constituents at those wavelengths. Again, it is possible to solve each equation independently provided that the spectrum of one constituent does not interfere with the spectrum of the other (i.e., the bands are well resolved).

Unfortunately, the equations above make the assumption that the absorbance at wavelength 1 is entirely due to constituent A and the absorbance at wavelength 2 is entirely due only to constituent B. As with the [Least Squares Regression](#) model, this requires finding two wavelengths in the training set of spectra that exclusively represent constituents A and B. With complex mixtures, or even simple mixtures of very similar materials, this is a difficult if not impossible task.

However, it is possible to get around this by taking advantage of another part of [Beer's Law](#); the



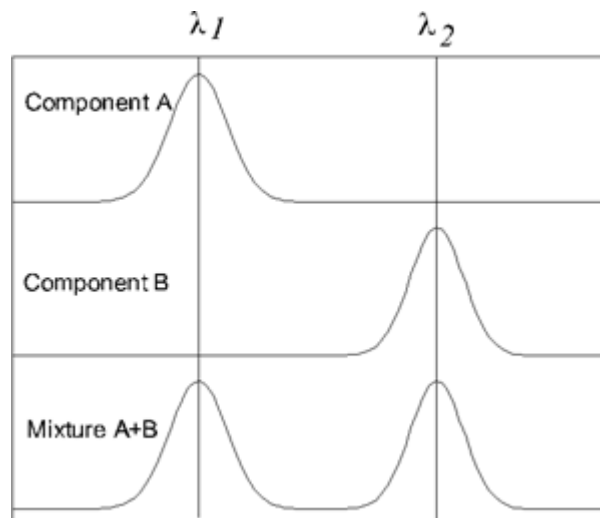
Hypothetical spectra of two alternative pure constituents, A and B, and a mixture of the two. In this case, the bands of the constituent spectra overlap, and the equations must be solved simultaneously for both A and B.

measurements or in the predictive ability of the model when the equations are used to predict unknowns. Once again, this never happens in the real world; there is always some amount of error. The existence of error is what requires running more than two samples in the first place.

It is necessary to amend the equations one more time to add a variable to compensate for the errors in the calculation of the absorbance:

$$A_{\lambda_1} = K_{a,\lambda_1}C_a + K_{b,\lambda_1}C_b + E_{\lambda_1}$$

$$A_{\lambda_2} = K_{a,\lambda_2}C_a + K_{b,\lambda_2}C_b + E_{\lambda_2}$$



Hypothetical spectra of two different pure constituents A and B and a mixture of the two. Since the constituent bands in the spectra do not overlap, the selected wavelengths could be used to solve separate equations for both A and B.

absorbances of multiple constituents at the same wavelength are additive. Thus, the two constituent equations for a single spectrum should really be:

$$A_{\lambda_1} = K_{a,\lambda_1}C_a + K_{b,\lambda_1}C_b$$

$$A_{\lambda_2} = K_{a,\lambda_2}C_a + K_{b,\lambda_2}C_b$$

All the equations presented so far assume that the calculated least squares line(s) that best fits the calibration samples is perfect. In other words, it has been assumed that there is no error in the

where  $E_{\lambda 1}$  and  $E_{\lambda 2}$  are the residual errors between the least squares fit line and the actual absorbances. When performing [Least Squares Regression](#), the "offset" coefficient (a.k.a., intercept, bias) performs the same function. In these terms, the  $E$  values can be thought of as the calibration offset or bias. It is obvious to see that this will always be zero when fitting only two points (i.e., only two calibration mixture samples). However, as with most calibration models, Classical Least Squares usually requires many more training samples to build an accurate calibration. As long as the same number (or more) wavelengths are used as there are constituents, it is possible to calibrate for all constituents simultaneously.

The next problem is how to solve all these equations. If you have ever tried to solve simultaneous equations by hand, you know this is a very tedious process. If more than 2 constituents are present or more than two wavelengths are used, it gets even harder. A particularly efficient way of solving simultaneous equations is to use linear algebra, also known as matrix mathematics. This technique still requires many calculations, but the rules are straightforward and are perfectly suited for computers. In matrix terms, the previous equation can be formulated as:

$$\begin{bmatrix} A_{\lambda 1} \\ A_{\lambda 2} \end{bmatrix} = \begin{bmatrix} K_{a,\lambda 1} & K_{b,\lambda 1} \\ K_{a,\lambda 2} & K_{b,\lambda 2} \end{bmatrix} \begin{bmatrix} C_a \\ C_b \end{bmatrix} + \begin{bmatrix} E_{\lambda 1} \\ E_{\lambda 2} \end{bmatrix}$$

or, more simply:

$$A = KC + E$$

In this case,  $A$  represents a (2 x 1) matrix of absorbances at the two selected wavelengths,  $K$  is a (2 x 2) matrix of the absorptivity constants,  $C$  is a (2 x 1) matrix of the concentrations of the two constituents, and  $E$  is the (2 x 1) matrix of absorbance error, or offset.

This model can be extended to performing calculations using many more wavelengths than just two. In fact, as long as the number of wavelengths used for the model is LARGER than the number of constituents in the mixtures, any number of wavelengths can be used. In fact, it is not unusual to use the entire spectrum when calibrating Classical Least Squares models. In this case the matrices look like:

$$\begin{bmatrix} A_{1,1} & \dots & A_{n,1} \\ \vdots & & \vdots \\ A_{1,p} & \dots & A_{n,p} \end{bmatrix} = \begin{bmatrix} K_{1,1} & \dots & K_{m,1} \\ \vdots & & \vdots \\ K_{1,p} & \dots & K_{m,p} \end{bmatrix} \begin{bmatrix} C_{1,1} & \dots & C_{1,n} \\ \vdots & & \vdots \\ C_{m,1} & \dots & C_{m,n} \end{bmatrix}$$

where  $A$  is a matrix of spectral absorbances,  $K$  is the matrix of absorptivity constants and  $C$  is the matrix of constituent concentrations. (The  $E$  matrix is not shown for space reasons, but has the same dimensionality as the  $A$  matrix.) The subscripts indicate the dimensionality of the matrix;  $n$  is the number of samples (spectra),  $p$  is the number of data points (wavelengths) used for calibration, and  $m$  is the number of constituents in the sample mixtures.

Using matrix algebra, it is trivial for a computer to solve these equations and produce the  $K$  matrix in

the above equation (the matrix of absorptivity coefficients). Just by the nature of matrix algebra, the solution gives the best fit least squares line(s) to the data. Once the equation is solved for the  $\mathbf{K}$  matrix, it can be used to predict concentrations of unknown samples.

For those familiar with linear algebra, to solve for the  $\mathbf{K}$  matrix requires computing the matrix equation:

$$\mathbf{K} = \mathbf{A} \mathbf{C}^{-1}$$

where  $\mathbf{C}^{-1}$  is the inverse of the constituent concentration matrix. Unfortunately, computing the inverse of a matrix requires that the matrix be square (having the same number of rows and columns). Unless the calibration set has exactly the same number of samples as constituents, this will not be true (remember, more samples are usually used to get the best representation of the true calibration equation).

This does not mean that the above equation cannot be solved. An alternative to computing the true inverse of the  $\mathbf{C}$  matrix is to compute its "pseudo-inverse", as follows:

$$\mathbf{K} = \mathbf{A} \mathbf{C}'(\mathbf{C} \mathbf{C}')^{-1}$$

where  $\mathbf{C}'$  is the matrix transpose (pivot the matrix so that the rows become the columns) of the constituent concentrations matrix.

This method has the advantage of being able to use large regions of the spectrum, or even the entire spectrum, for calibration to gain an averaging effect for the predictive accuracy of the final model. One interesting side effect is that if the entire spectrum is used for calibration, the rows of the  $\mathbf{K}$  matrix are actually spectra of the absorptivities for each of the constituents. These will actually look very similar to the pure constituent spectra.

However, this technique does have one major disadvantage: the equations must be calibrated for every constituent in the mixtures. Otherwise, the ignored constituents will interfere with the analysis and give incorrect results. This means that the complete composition of every calibration sample must be known, and that predicted "unknowns" must be mixtures of exactly the same constituents.

This limitation of the CLS model can be more easily understood by taking a closer look at the model equation:

$$A_{11} = K_{11}C_{11} + K_{12}C_{21} + E_{11}$$

Notice that the absorbance at a particular wavelength is calculated from the sum of all the constituent concentrations multiplied by their absorptivity coefficients. If the concentration of any constituent in the sample is omitted, the predicted absorbance will be incorrect. This means that the CLS technique can only be applied to systems where the concentration of every constituent in the sample is known. If the mixture is complex, or there is the possibility of contaminants in the "unknown" samples that were not present in the calibration mixtures, then the model will not be able to predict the constituent concentrations accurately.

### CLS Advantages

- Based on [Beer's Law](#).
- Calculations are relatively fast.
- Can be used for moderately complex mixtures.
- Calibrations do not necessarily require wavelength selection. As long as the number of wavelengths exceeds the number of constituents, any number (up to the entire spectrum) can be used.
- Using a large number of wavelengths tends to give an averaging effect to the solution, making it less susceptible to noise in the spectra.

### CLS Disadvantages

- Requires knowing the complete composition (concentration of every constituent) of the calibration mixtures.
- Not useful for mixtures with constituents that interact.
- Very susceptible to baseline effects since equations assume the response at a wavelength is due entirely to the calibrated constituents.

## Inverse Least Squares

One of the most widely used spectroscopic quantitation methods is Inverse Least Squares, also known as **Multiple Linear Regression** and **P-Matrix**. Most methods based on [Beer's Law](#) assume that there is little or no interference in the spectrum between the individual sample constituents or that the concentrations of all the constituents in the samples are known ahead of time. In real world samples, it is very unusual, if not entirely impossible to know the entire composition of a mixture sample. Sometimes, only the quantities of a few constituents in very complex mixtures of multiple constituents are of interest. Obviously, these simpler calibration methods will fail for these samples since the full compositional chemistry and pure spectroscopic composition is not known. One solution to this problem is to take advantage of algebra to rearrange Beer's Law and express it as:

$$C = \frac{A_{\lambda}}{\epsilon_{\lambda} b}$$

where the same relationships between the spectral response at a single wavelength ( $A_{\lambda}$ ), the constituent absorptivity constant ( $\epsilon_{\lambda}$ ), the pathlength of light ( $b$ ) and the constituent concentration ( $C$ ) are maintained. By combining the absorptivity coefficient ( $\epsilon_{\lambda}$ ) and the pathlength ( $b$ ) into a single constant, this can also be expressed as:

$$C = P A_{\lambda} + E$$

where  $E$  is a matrix of concentration prediction error.

Due to the powers of mathematics, this seemingly trivial variation has tremendous implications for the experiment. This expression of Beer's Law implies that the concentration is a function of the absorbances at a series of given wavelengths. This is entirely different from Classical Least Squares

(CLS), where absorbance at a single wavelength is calculated as an additive function of the constituent concentrations. Consider the following two equations:

$$\begin{aligned}C_a &= A_{a1}P_{a,1} + A_{a2}P_{a,2} + E_a \\C_b &= A_{b1}P_{b,1} + A_{b2}P_{b,2} + E_b\end{aligned}$$

Notice that even if the concentrations of all the other constituents in the mixture are not known, the matrix of coefficients ( $\mathbf{P}$ ) can still be calculated correctly. This model, known by many different names including Inverse Least Squares (ILS), Multiple Linear Regression (MLR), or P-Matrix seems to be the best approach for almost all quantitative analyses since no knowledge of the sample composition is needed beyond the concentrations of the constituents of interest.

The selected wavelengths must be in a region where there is a contribution of that constituent to the overall spectrum. In addition, measurements of the absorbances at different wavelengths are needed for each constituent. In fact, in order to accurately calibrate the model, measurements of at least one different wavelength is needed for each additional independent variation (constituent) in the spectrum.

Again for those interested in matrix algebra, the  $\mathbf{P}$  matrix of coefficients can be solved by computing:

$$\mathbf{P} = \mathbf{C} \mathbf{A}^{-1}$$

but as with CLS before, if the A matrix is not square, the pseudo-inverse must be used instead:

$$\mathbf{P} = \mathbf{C} \mathbf{A}'(\mathbf{A} \mathbf{A}')^{-1}$$

This model seems to give the best of all worlds. It can accurately build models for complex mixtures when only some of the constituent concentrations are known. The only requirement is selecting wavelengths that correspond to the absorbances of the desired constituents.

Unfortunately, the ILS calibration approach does have some drawbacks. Due to the dimensionality of the matrix equations, the number of selected wavelengths cannot exceed the number of training samples. In theory, it should be possible to just measure many more training samples to allow for additional wavelengths, but this causes a new problem. The absorbances in a spectrum tend to all increase and decrease together as the concentrations of the constituents in the mixture change. This effect, known as collinearity, causes the mathematical solution to become less stable with respect to each constituent.

Another problem with adding more wavelengths to the model, is an effect known as overfitting. Generally, starting from very few wavelengths, and adding more to the model (provided they are chosen to reflect the constituents of interest) will improve the prediction accuracy. However, at some point, the predictions will start to get worse. When the number of wavelengths increases in the calibration equations, the likelihood that "unknown" samples will vary in exactly the same manner decreases. When too much information in the spectrum is used to calibrate, the model starts to include the spectral noise which is unique to the training set and the prediction accuracy for unknown samples suffers.

In ILS, the averaging effect gained by selecting many wavelengths in the CLS method is effectively lost. Therefore wavelength selection is critically important to building an accurate ILS model. Ideally, there is a crossover point between selecting enough wavelengths to compute an accurate least squares line and selecting few enough so that the calibration is not overly affected by the collinearity of the spectral data.

Many of today's software packages that perform ILS (MLR) calibrations use sophisticated algorithms to find the "best" set of wavelengths to use for each individual constituent of interest. They attempt to search through the wavelengths and try different combinations to locate that cross-over point. Luckily, the calculations involved in computing the P matrix are very fast. However, when you consider that a spectrum may have as many as 2000 wavelength data points, it's obvious to see that calculating ILS models for all possible combinations of wavelengths can be an excruciating task.

Inverse Least Squares is an example of a multivariate method. In this type of model, the dependent variable (concentration) is solved by calculating a solution from multiple independent variables (in this case, the responses at the selected wavelengths). It is not possible to work backwards from the concentration value to the independent spectral response values as an infinite number of possible solutions exist. However, the main advantage of a multivariate method is the ability to calibrate for a constituent of interest without having to account for any interferences in the spectra.

Since it is not necessary to know the composition of the training mixtures beyond the constituents of interest, the ILS method is better suited to more complex types of analyses not handled by the CLS approach. It has been used for samples ranging from natural products (such as wheat, wool, cotton and gasoline) to manufactured products.

### **ILS Advantages**

- Based on Beer's Law.
- Calculations are relatively fast.
- Multivariate model allows calibration of very complex mixtures since only knowledge of constituents of interest is required.

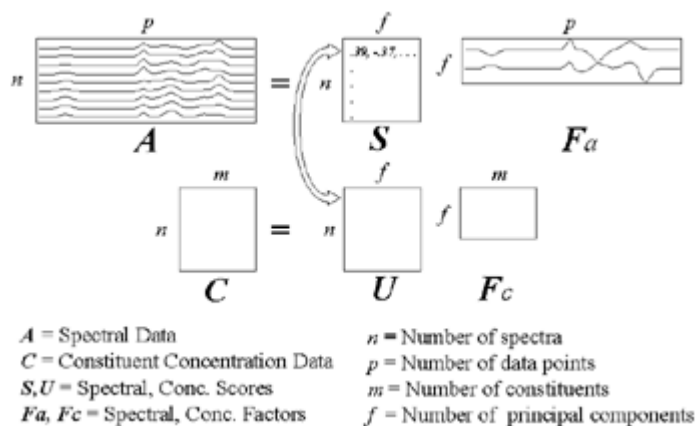
### **ILS Disadvantages**

- Wavelength selection can be difficult and time consuming. Must avoid collinear wavelengths
- Number of wavelengths used in the model limited by the number of calibration samples.
- Generally, a large number of samples are required for accurate calibration.
- Collecting calibration samples and measuring via a primary calibration can be difficult and tedious.

## **Partial Least Squares**

Partial Least Squares (PLS) is a quantitative spectral decomposition technique that is closely related to Principal Component Regression (PCR). However, in PLS, the decomposition is performed in a slightly different fashion. Instead of first decomposing the spectral matrix into a set of eigenvectors and scores, and regressing them against the concentrations as a separate step, PLS actually uses the concentration information during the decomposition process. This causes spectra containing higher constituent concentrations to be weighted more heavily than those with low concentrations. Thus, the eigenvectors and scores calculated using PLS are quite different from those of PCR. The main idea of PLS is to get as much concentration information as possible into the first few loading vectors.

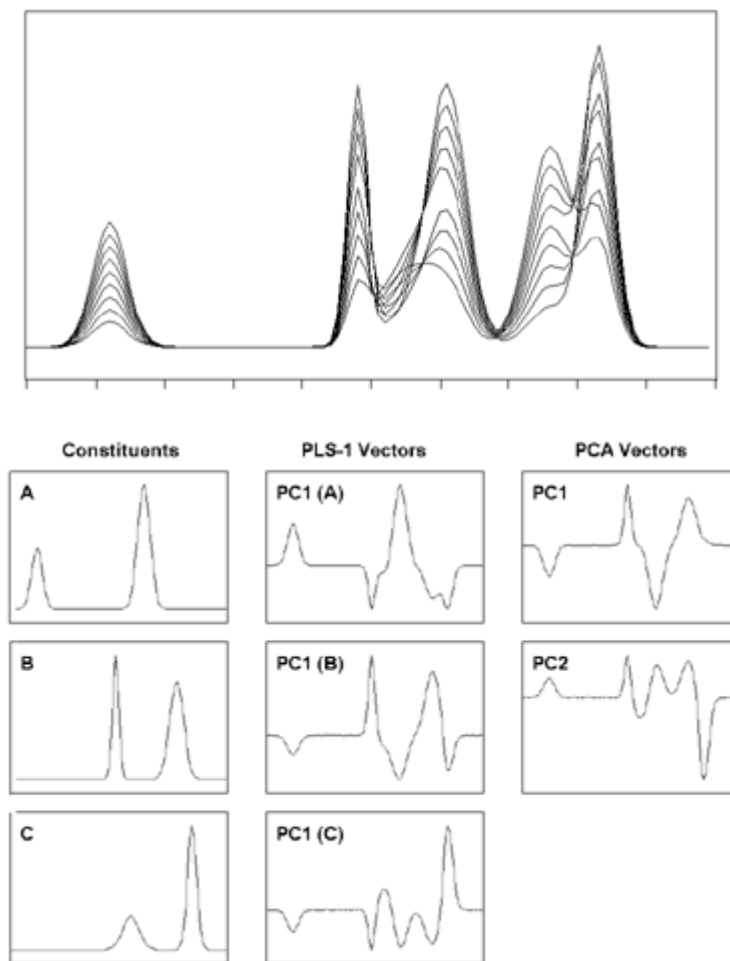
In actuality, PLS is simply taking advantage of the correlation relationship that already exists between the spectral data and the constituent concentrations. Since the spectral data can be decomposed into its most common variations, so can the concentration data! In effect, this generates two sets of vectors and two sets of corresponding scores; one set for the spectral data, and the other for the constituent concentrations. Presumably, the two sets of scores are related to each other through some type of regression, and a calibration model is constructed.



PLS is similar to PCA/PCR. However, in PLS the constituent concentration data is included in the decomposition process. In fact, both the spectral and concentration data are decomposed simultaneously, and the scores ( $S$  and  $U$ ) are "exchanged" as each new factor is added to the model.

Now this is actually a bit of an over-simplification. Unlike PCR, PLS is a one step process. In other words, there is no separate regression step. Instead, PLS performs the decomposition on both the spectral and concentration data simultaneously. As each new factor is calculated for the model, the scores are "swapped" before the contribution of the factor is removed from the raw data. The newly reduced data matrices are then used to calculate the next factor, and the process is repeated until the desired number of factors is calculated. Unfortunately this makes the model equations for PLS significantly more complex than those of PCR. For those who are interested, the algorithms for calculating the PLS model eigenvectors and scores are shown in a later section.

As mentioned previously, one of the main advantages of PLS is that the resulting spectral vectors are directly related to the constituents of interest. This is entirely unlike PCR, where the vectors merely represent the most common spectral variations in the data, completely ignoring their relation to the constituents of interest until the final regression step.



The vectors generated by PLS (especially PLS-1) are more directly related to the constituents of interest than those from PCA. The left column shows the spectra of the "pure" constituents used to construct the data set in the Top image. The center column shows the first PLS-1 vector for each constituent calculated from the data set, while the right column shows the first two PCA vectors for the same data.

There are actually two versions of the PLS algorithm; PLS-1 and PLS-2. The differences between these methods are subtle but have very important effects on the results. Like the PCR method, PLS-2 calibrates for all constituents simultaneously. In other words, the results of the spectral decomposition for both of these techniques give one set of scores and one set of eigenvectors for calibration. Therefore, the calculated vectors are not optimized for each individual constituent. This may sacrifice some accuracy in the predictions of the constituent concentrations, especially for complex sample mixtures. In PLS-1, a separate set of scores and loading vectors is calculated for each constituent of interest. In this case, the separate sets of eigenvectors and scores are specifically tuned for each constituent, and therefore, should give more accurate predictions than PCR or PLS-2.

There is, however, a minor disadvantage in using the PLS-1 technique: the speed of calculation. Since a separate set of eigenvectors and scores must be generated for every constituent of interest, the calculations will take more time. For training sets with a large number of samples and constituents, the

increased time of calculation can be significant.

PLS-1 may have the largest advantage when analyzing systems that have constituent concentrations that are widely varied. For example, a set of calibration spectra contains one constituent in the concentration range of 50 to 70% and a second constituent in the range of 0.1 to 0.5%. In this case, PLS-1 will almost certainly predict better than the other techniques. If the concentration ranges of the constituents are approximately the same, PLS-1 may have less of an advantage over PLS-2 and will definitely take longer to calculate.

### PLS Advantages

- Combines the full spectral coverage of [CLS](#) with partial composition regression of [ILS](#).
- Single step decomposition and regression; eigenvectors are directly related to constituents of interest rather than largest common spectral variations.
- Calibrations are generally more robust provided that calibration set accurately reflects range of variability expected in unknown samples.
- Can be used for very complex mixtures since only knowledge of constituents of interest is required.
- Can sometimes be used to predict samples with constituents (contaminants) not present in the original calibration mixtures.

While all of these techniques have been successfully applied for spectral quantitative analysis, the arguments in the literature generally show that PLS has superior predictive ability. In most cases, PLS methods will give better results than PCR, and PLS-1 will be more accurate than PLS-2. However, there are many documented cases in the literature where certain calibrations have performed better by using PCR or PLS-2 instead of PLS-1. Unfortunately, there are no definite rules, and only good research practices can determine the best model for each individual system.

### PLS Disadvantages

- Calculations are slower than most Classical methods, especially PLS-1.
- Models are more abstract, thus more difficult to understand and interpret.
- Generally, a large number of samples are required for accurate calibration.
- Collecting calibration samples can be difficult; must avoid collinear constituent concentrations.

### Calculating PLS Eigenvectors and Scores

This section is for those who are interested in knowing the mechanics of the PLS calculation. As mentioned above, there are two variants of this algorithm known as PLS-1 and PLS-2. In fact, PLS-1 is a reduced subset of the full PLS-2. The algorithms have been combined here, with appropriate notes on where they differ. Note that a PLS-2 model of a training set with only one constituent is identical to a PLS-1 model for the same data.

The main difference between PLS and PCR is that the concentration information is included in the calculations during the spectral decomposition. This results in two sets of eigenvectors; a set of spectral "loadings" ( $Bx$ ) which represent the common variations in the spectral data, and a set of spectral "weights" ( $W$ ) which represent the changes in the spectra that correspond to the regression constituents. Correspondingly, there are two sets of scores: one for the spectral data ( $S$ ) and another for the concentration data ( $U$ ).

The following description assumes that the matrices involved have the following dimensions:  $A$  is an  $n$

by  $\mathbf{p}$  matrix of spectral absorbances,  $\mathbf{C}$  is an  $n$  by  $m$  matrix of constituent concentrations,  $\mathbf{S}$  is an  $f$  by  $n$  matrix of spectral scores,  $\mathbf{U}$  is an  $f$  by  $n$  matrix of concentration weighted scores,  $\mathbf{Bx}$  is an  $f$  by  $p$  matrix of spectral loading vectors,  $\mathbf{W}$  is an  $f$  by  $p$  matrix of spectral weighting vectors,  $\mathbf{By}$  is an  $f$  by  $m$  matrix of the constituent loading vectors and  $\mathbf{V}$  is a  $1$  by  $f$  vector of the PLS model cross products. In this case,  $n$  is the number of samples (spectra),  $p$  is the number of data points (wavelengths),  $m$  is the number of constituents, and  $f$  is the number PLS eigenvectors. When used, the subscripts on the matrices indicate a matrix row.

1. Set the weighting scores to a starting value:  $U_i = C'_i$   
(For PLS-1 or single constituent models, use the desired constituent vector. For PLS-2, use the first constituent column vector.)
2. Calculate the spectral weighting vector:  $W_i = U'_i A$
3. Normalize the weighting vector to unit length:  $W_i = W_i / (W_i W'_i)$
4. Calculate the spectral scores:  $S_i = A W_i$   
(For PLS-1 or single constituent model, set  $By_i = 1$  and skip to step 9.)
5. Calculate the concentration loading vector:  $By_i = S_i C$
6. Normalize the concentration loading vector to unit length:  $By_i = By_i / (By_i By'_i)$
7. Calculate new weighting scores:  $U_i = By_i C'$
8. Check for convergence by comparing new  $U_i$  scores to the previous pass for this vector. If this is the first pass for the current vector, or the scores are not the same, go back to step 2. If the scores are effectively the same, continue with step 9.
9. Calculate the PLS cross product for this vector:  $V_i = S_i U'_i / (S_i S'_i)$
10. Calculate the spectral loading vector:  $Bx_i = S_i A$
11. Normalize the spectral loading vector by the spectral scores:  $Bx_i = Bx_i / (S_i S'_i)$
12. Remove contribution of the vector from the spectral data:  $A = A - S'_i Bx_i$
13. Remove contribution of the vector from the concentration data:  $C = C - (S'_i By_i) V_i$
14. Increase vector counter,  $i = i + 1$  and go back to step 1. Continue until all desired factors are calculated ( $i = f$ ).
15. If performing PLS-1, reset  $A$  back to the original training set values and redo all steps using a different constituent in step 1. Note that this generates a completely different set of  $S$ ,  $U$ ,  $W$ ,  $Bx$ ,  $By$  and  $V$  matrices for every constituent!

### Predicting Samples with a PLS Model

The following are the calculational steps used to predict a spectrum against a PLS model. The variable descriptions are as above, except that  $\mathbf{Au}$  is the  $1$  by  $p$  vector of the spectral responses of the sample being predicted, and  $\mathbf{Cu}$  is the  $1$  by  $m$  vector of the predicted constituent concentrations. Initially,  $\mathbf{Cu}$  is set to zero, and the vector counter  $i$  to one.

1. Calculate the unknown spectral score for a weighting vector:  $S_i = W_i' Au$
2. Calculate the concentration contribution for the vector:  $Cu = Cu + (Byi' Si Vi)$
3. Remove the spectral contribution of the vector:  $Au = Au - (Si Bxi')$
4. Increment the vector counter  $i = i + 1$  and go back to step 1. Continue until all desired factors are calculated ( $i = f$ ).
5. If performing PLS1, reset the data in  $Au$  back to the original unknown spectrum values and repeat from step 1 with the next set of constituent vectors.

Note that the data remaining in the  $Au$  vector after all factors have been removed is the residual spectrum.

## Discriminant Analysis

There are many advantages of using spectroscopy as a detection technique for quality control of complex samples. It is fast, requires little or no sample preparation for most types of samples, and can be implemented at or near the source of the samples. However, many times, quantitative methods are employed to simply gauge the suitability of the material being measured. In a significant number of cases, the only result that is desired is to know whether the sample falls within a defined range of allowed variability to determine if the material is of the desired quality. It is not always necessary to measure the quantities of the constituents in the sample to meet this goal.

Multivariate quantitative models such as Principal Component Regression ([PCR](#)) and Partial Least Squares ([PLS](#)) generally require a large number of training samples to build accurate calibrations. In turn, this requires a lot of initial work collecting all the samples and measuring the concentrations of the constituents by the primary method before the data can even be used for model building. The accuracy of the calibration is limited by the accuracy of the primary method used to get the concentration values. If the primary method is not very good, the multivariate model will not be very good either. If merely knowing that a sample is of a given quality is required and the quantity of the constituents is not needed, using a quantitative model adds a lot of extra work to simply determine if the sample is the same as the training set data.

In addition, the quantities of the constituents are usually not the whole story when measuring product quality. Sometimes samples can be contaminated with other compounds and impurities. Generally, quantitative models will always predict reasonable values for the calibrated constituents, provided the spectra of the unknown samples are fairly similar to the training set. But the reported concentrations alone will not indicate if the samples are contaminated.

In some cases, the constituent information is simply not available for the samples to be calibrated.

There may not be a primary calibration method available for the constituent(s) of interest, or the samples may be simply too complex. Another very likely possibility is that it is possible to collect all the primary constituent information, but the work involved to actually do it would be prohibitively expensive. However, the spectrum of a sample is unique to the composition of its constituents. Samples of the same or similar composition quality should have spectra that are very similar as well. Theoretically, it should be possible to tell the difference between a "good" sample and a "bad" one by only comparing their spectra.

Unfortunately, the tolerances required for determining the differences between spectra in quality control applications cannot usually be met by simple methods such as visual inspection or [spectral](#)

subtraction. In addition to requiring user interaction (and they are therefore subjective methods inappropriate for quality control), they cannot be easily used by anyone other than a trained spectroscopist. What is needed instead is an unambiguous mathematical method for spectral matching. What has been described here is the basis of discriminant analysis. The primary purpose is to classify samples into well defined groups or categories based on a training set of similar samples without prior or with limited knowledge of the composition of the group samples. The ultimate aim of discriminant analysis is to unambiguously determine the identity or quality of an unknown sample. A good discriminant algorithm is one that can "learn" what the spectrum of a sample looks like by "training" it with spectra of the same material. For this reason, discriminant analysis is sometimes called pattern recognition.

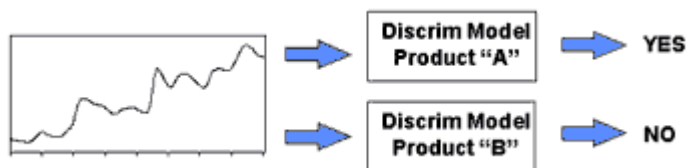
There are two basic applications for spectroscopic discriminant analysis: sample purity/quality and sample identification/screening. In the capacity of sample quality checking, discriminant analysis methods can replace many quantitative methods currently used. In effect, the algorithm gives an indication of whether the spectrum of the "unknown" sample matches the spectra from samples taken previously that were known to be of "good" quality. Some algorithms can even give statistical measurements of the quality of the match.



Quality control/assurance application of spectroscopic discriminant analysis. The spectrum of the sample is compared against the model to determine if it matches the training data for the model. If the training set was constructed from spectra of samples that were of known quality, the model can accurately predict if the sample is of the same quality by matching the spectrum and giving a "yes" or "no" answer.

When discriminant analysis is used in a product identification or product screening mode, the spectrum of the "unknown" is compared against multiple models. The algorithm will give an indication of the likelihood of the spectrum matching a model and the product can then be identified as a particular material. This mode of discriminant analysis is sometimes used for grading materials as well. For this application, each model is built from a set of samples that represent a particular grade/purity/quality of the material. When the unknown spectrum is predicted against the models, the material is classified as the closest match (or no match at all).

The analyst can control how the discrimination is calculated. Any samples in the training set become representative of the allowed form of the spectrum. For example, discriminant analysis could be also used to classify samples into chemical classes by making training sets of spectra of different compounds that share similar functional groups. As long as enough samples are used to represent the range of variability found in those types of compounds, "unknowns" could be chemically classified by comparing them to all the training sets and looking for a match.



Sample identification/screening application of spectroscopic discriminant analysis. The spectrum of the sample is compared to multiple models of different materials or different levels of quality of the same material. The models can predict the likelihood that the sample matches the training spectra they were constructed from, again giving a "yes" or "no" answer.

There are a vast number of useful analyses that can be solved by discriminant analysis. The main advantage these methods have is that they are generally easier to apply to spectroscopic analysis than quantitative methods since they do not need any primary calibration data to build the model. They give simple "pass" or "fail" answers as to how well the samples match by comparing them to training sets of the desired quality samples. They learn to recognize the spectra of materials based entirely on the spectral data itself without any other external information other than the analyst's logical grouping of the spectra into training sets.

Many different methods have been developed for performing discriminant analysis on spectra. One class of algorithm that is already familiar to many spectroscopists is [Spectral Library Searching using Euclidean Distance](#). In these algorithms, the spectrum of an unknown sample is compared against many different spectra in a library of known compounds. By comparing the responses at all wavelengths in the "unknown" spectrum to the corresponding responses in a series of known (or "library") spectra, a list of the closest matches can be identified by ranking the known spectra by a calculated "Hit Quality Index".

Many commercially available library search programs use these techniques to generate a list of the most likely matches of the unknown sample. However, there are many problems with this technique. First, search techniques simply identify samples as the materials from the closest matching spectrum in the library. If the library does not contain any spectra of the "true" compound, it will just report the best match it found regardless of whether it is really even the same class of material.

In addition, the spectral library search technique is only sensitive to the general spectral shapes and patterns, and not to very subtle variations within the sample. If the variations in between the spectra of a "good" sample and "bad" sample cannot be easily seen by visual inspection, chances are a spectral library search will not be able to do it either. Typically Spectral Library Search algorithms cannot be trained to recognize a range of variability in the data since the spectrum of the unknown is only compared to a single representative spectrum for each different class of material.

Another problem is that the spectra must have flat baselines in order for these methods to work properly. As seen earlier in the discussions of preprocessing methods for quantitative spectroscopy, there are methods for accomplishing this with little or no user interaction. However, these methods are very sensitive to baseline instabilities, and the correction applied must be very good to have any degree of success with these methods.

Finally, the "Hit Quality Index" does not provide any absolute measure of the probability that the sample actually is the same as the library sample. The arbitrary scale of the Hit Quality values (0 to 1) does not give a very good statistical measure of the similarity of the spectra. In short, using only a single training spectrum to represent all possible samples in the future does not give the analyst any statistical assurance that the spectra are truly the same or different. It only provides a relative measure for all the library samples. For anyone who has tried simple library search techniques for spectrally similar samples, this result is all too obvious.

There have been many other methods put forth in the literature including K-nearest neighbor, Cluster analysis, PCA Factorial Discriminant Analysis, SIMCA, BEAST, and others (see [References](#)).

Many of these methods use [Principal Component Analysis](#) as a spectral data compression technique. PCA decomposes a set of spectra into their most common variations (factors) and produces a small set

of well defined numbers (scores) for each sample that represent the amount of each variation present in the spectrum. Similar to using the way [PCR](#) and [PLS](#) quantitation methods use these scores for creating a calibration equation, they can also be used for discrimination since they provide an accurate description of the entire set of training spectra. However, many of the methods listed above only utilize the first few significant factors for discrimination. In many cases, only the first two factors are used. Thus, only a limited portion of the total spectral information available for a class of material is actually used; the rest is simply discarded.

## Principal Component Regression

The Principal Component Regression method combines the Principal Component Analysis ([PCA](#)) spectral decomposition with an Inverse Least Squares ([ILS](#)) regression method to create a quantitative model for complex samples. Unlike quantitation methods based directly on [Beer's Law](#) which attempt to calculate the absorptivity coefficients for the constituents of interest from a direct regression of the constituent concentrations onto the spectroscopic responses, the PCR method regresses the concentrations on the PCA scores.

The eigenvectors of a PCA decomposition represent the spectral variations that are common to all of the spectroscopic calibration data. Therefore, using that information to calculate a regression equation (in place of the straight spectral responses) will produce a robust model for predicting concentrations of the desired constituents in very complex samples. It is interesting to note that the PCA factors matrix,  $F$ , performs a similar task to the  $K$  matrix in the Classical Least Squares ([CLS](#)) model; it stores the "constituent" spectral data. This does not mean that the rows of the  $F$  matrix are the spectra of the pure constituents; because they are not. However, they cannot be used alone without the scores matrix  $S$  to represent the original data (as in CLS, it needs the  $C$  matrix to perform the same function).

On the other hand, the scores in the  $S$  matrix are unique to each calibration spectrum, and just as a spectrum is represented by a collection of absorbances at a series of wavelengths, it can also be series of scores for a given set of factors. Much like the classical models performed a regression of the concentration  $C$  matrix directly on the spectral absorbances in the  $A$  matrix, it is also possible to regress  $C$  against the scores  $S$  matrix.

In this case, the regression technique from the [ILS](#) model is obviously the best choice. This gives the model the best qualities of the ILS method, such as, no a priori knowledge of the complete sample composition and some robustness in predictions with respect to contaminant constituents not present in the original calibration mixtures. The model equation is therefore:

$$C = B S + E_c$$

where  $C$  is the  $m$  by  $n$  matrix of constituent concentrations,  $B$  is an  $m$  by  $f$  matrix of the regression coefficients and the  $S$  matrix is the scores from the PCA model. The dimensions of the matrices are  $n$  for the number of samples (spectra),  $m$  for the number of constituents used for calibration, and  $f$  for the number PCA eigenvectors. As with ILS, the  $B$  coefficients matrix can be solved by the regression:

$$B = C S' (S S')^{-1}$$

Thus the name for this type of model is Principal Components Regression; it combines Principal Component Analysis and Inverse Least Squares Regression to solve the calibration equation for the model. All that remains is to come up with a single unified equation that represents the PCR model. Therefore, rearranging the matrix model equation from before to represent the scores as a function of the spectral absorbances and the eigenvectors produces:

$$S = A F'$$

It is not necessary to use the inverse (or pseudo inverse) of  $F$  to solve this equation. This is due to the fact that when PCA is used to solve the spectral model, the resulting  $F$  matrix of eigenvectors is a special type of matrix called an orthonormal matrix. This type of matrix has a very interesting quality: when the matrix is multiplied by its own transpose, the identity matrix is the result. Multiplying any matrix by the identity matrix is the same as multiplying a single number by one; the result is always the number again. So, to get the equation for the scores, both sides of the earlier PCA model equation for the spectral data were simply multiplied by the transpose of the  $F$  matrix. Finally, by combining the concentration equation with the scores equation, the final PCR model equation emerges:

$$C = B A F' + E_c$$

where  $C$  is the  $m$  by  $n$  matrix of constituent concentrations,  $B$  is an  $m$  by  $f$  matrix of the regression coefficients,  $A$  is an  $n$  by  $p$  matrix of spectral absorbances, and  $F$  is an  $f$  by  $p$  matrix of eigenvectors. The dimensions of the matrices are  $n$  for the number of samples (spectra),  $m$  for the number of constituents used for calibration,  $p$  for the number of data points (wavelengths) used for calibration, and  $f$  for the number PCA eigenvectors.

The PCR calibration model is not completely free of problems, however. It is important to note that PCR is a two-step process; the PCA eigenvectors and scores are calculated and then the scores are regressed against the constituent concentrations using a regression method similar to ILS. Remember that the ILS method can build accurate calibrations, provided that the selected variables (in the earlier discussion, the variables were the responses at selected wavelengths) are physically related to the properties (constituent concentrations) they are regressed against. However, the PCA factors/scores are calculated independently of any knowledge of these concentrations. They merely represent the largest common variations among all the spectra in the training set. Presumably, these variations will be mostly related to changes in the constituent concentrations, but there is no guarantee this will be true. In fact, many PCR models include more factors than are actually necessary as some of the eigenvectors are not related to any of the constituents of interest. Ideally, a PCR model should be built by performing a selection on the scores (much like a selection of wavelengths for an ILS model) to determine which factors should be used to build a model for each constituent. In practice, this is a difficult process, both in terms of developing the selection rules and making it simple to perform. Most commercial chemometric software packages do not support this type of PCR model. In addition, like the ILS method, the predictive ability of the PCR model will suffer if the constituent concentrations are collinear. Again, this means that a relatively large number of the training set samples are required, and they must be tested by the primary calibration method to determine the "randomness" of the constituent concentrations.

### PCR Advantages

- Does not require wavelength selection. Any number can be used; usually the whole spectrum, or large

regions.

- Larger number of wavelengths gives averaging effect, making model less susceptible to spectral noise.
- PCA data compression allows using inverse regression to calculate model coefficients; can calibrate only for constituents of interest.
- Can be used for very complex mixtures since only knowledge of constituents of interest is required.
- Can sometimes be used to predict samples with constituents (contaminants) not present in the original calibration mixtures.

### **PCR Disadvantages**

- Calculations are slower than most Classical methods.
- Optimization requires some knowledge of PCA; models are more complex to understand and interpret.
- No guarantee PCA vectors directly correspond to constituents of interest.
- Generally, a large number of samples are required for accurate calibration.
- Collecting calibration samples can be difficult; must avoid collinear constituent concentrations.

## **Chemometrics, Preprocessing Techniques**

[Light Scattering Correction, Multiplicative Scatter Correction\(MSC\)](#)

[Light Scattering Correction, Standard Normal Variate \(SNV\) Transformation and Detrending](#)

[Correcting Sample Pathlength Differences](#)

[Measured Pathlength Calibration](#)

[Sample Thickness Correction](#)

[Unit Area Normalization](#)

[Correcting Baseline Effects](#)

[First and Second Derivatives](#)

[Data Enhancement](#)

### **Correcting Sample Pathlength Differences**

[Beer's Law](#) states that there is a direct and linear relationship between sample concentration, pathlength and the absorbance of light at a particular wavelength. Most chemometric models are built using samples that vary in concentration, but the pathlength is fixed (except for diffuse reflectance measurements, however, MSC or SNV is used to correct this type of data). Factor-based chemometric models are not limited by this requirement, although the performance of the models is certainly better when this is true.

Unfortunately, it is not always possible to collect spectra of either training samples or "unknown" samples with a constant pathlength. For example, when measuring transmission spectra of thin films, it is very difficult to extrude polymers to a constant thickness every time. Obviously, if the pathlength varies in the sample, this will appear in the spectra as changes in response that are not correlated to concentration. Sometimes the factor based models can correct for these effects if the range of different pathlengths is not too large. However, the model will most certainly work much better if the pathlength effect can be removed altogether.

In addition to the methods listed here, MSC has also been applied to correct spectra of samples with indeterminate pathlengths, but not necessarily measured by diffuse reflectance. While it will not be as effective as either of the two specific pathlength corrections below ([Measured](#) and [Thickness](#)), it will

usually give better results than no correction at all. Some success has also been shown in using MSC to correct the pathlength effects in spectra measured using Attenuated Total Reflectance ([ATR](#)).

### Measured Pathlength Calibration

In some cases, it is possible to actually measure the exact pathlengths of the training samples. This information can be used during the model building step by including the pathlength as an extra constituent in the calibration. During the calibration calculations, all the concentration data is scaled to the entered pathlength for each sample. When an "unknown" sample is predicted, its pathlength is predicted at the same time. The concentrations are then un-scaled by the predicted pathlength before they are reported.

This method is useful in cases where the pathlengths of the training samples are easy to acquire. This can be used to correct for samples where the pathlength of the "unknowns" either cannot be (easily) measured or is expected to vary substantially over a period of time. The downside to this method is that accurately measuring the pathlengths of every training sample can be a tedious process.

### Sample Thickness Correction

This type of pathlength correction is sometime called the "internal standard" method. It is primarily used for samples that cannot be corrected by the [Measured Pathlength](#) method. One requirement for this method is that there must be an isolated band in every spectrum that arises from a constituent that does not vary in concentration in all samples, for both the training set and "unknowns" for prediction. Since the chosen spectral band is assumed to be concentration invariant in all samples, an increase or decrease of the absorption of that band in the spectrum can be assumed to be entirely due to an increase or decrease in the sample pathlength. Therefore, by normalizing the entire spectrum to the intensity of the band, the pathlength variation is effectively removed. The intensity of the band can be calculated as either the response at a single wavelength in the band (usually the peak maximum) or the integrated area.

One potential problem with this method is that it is extremely susceptible to baseline offset and slope effects in the spectrum. When calculating the pathlength normalization constants, the spectra must either be baseline corrected before creating the training set, or a local baseline must be calculated "on-the-fly" under the thickness correction band for each spectrum. The latter approach is usually recommended as the former method requires separate manual baseline correction of each "unknown" spectrum before prediction against the calibration model.

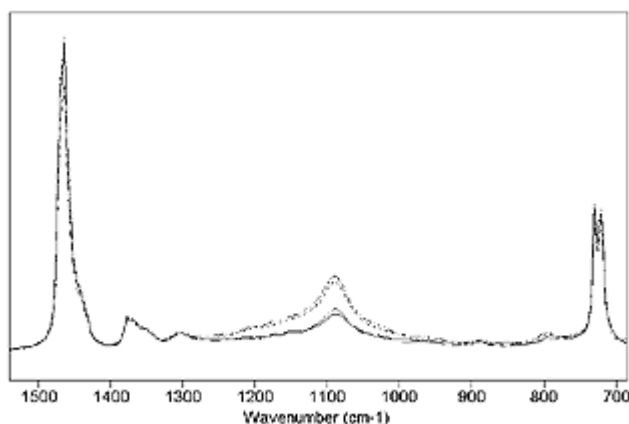


Figure 1. Spectra of 4 polymer thin films. The constituent of

interest is clearly the band in the middle: 2% (solid), 4% (dashed), 8% (dotted) and 10% (dash-dot). Note that the relative intensities of the constituent bands are not correct due to the varying thickness of the samples.

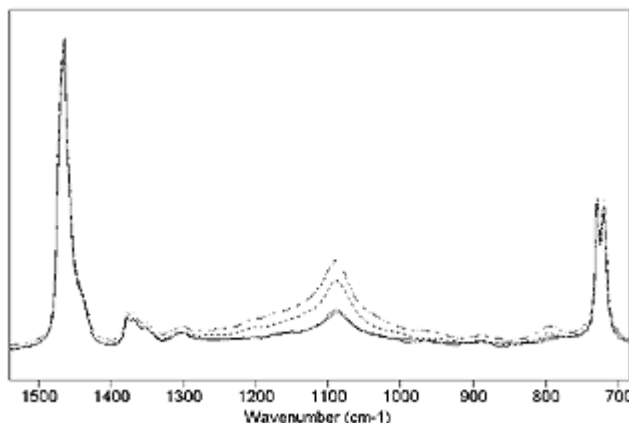


Figure 2. Spectra from Figure 5 after thickness correction. The integrated area of the band on the left (1525 - 1400  $\text{cm}^{-1}$ ) was used as the thickness correction factor. Notice that the relative intensities of the constituent bands now appear more in line with the known concentrations.

Thickness Correction is very useful for spectral measurements of samples that the pathlength cannot be guaranteed to be constant. However, it does require that the samples have a constant concentration constituent and that an isolated spectral band can be identified which is solely due to that constituent. Another use for Thickness Correction is to allow the calibration model to be pathlength independent. Instead of normalizing to a small region (isolated band) in the spectrum, a much larger region or even the entire spectral range is used to calculate the integrated area. This allows correcting samples with nearly any pathlength. The only requirement in using Thickness Correction in this manner is that the range of constituent concentrations must be relatively small. Large variations in the concentrations will cause the integrated areas of the spectra to vary mostly by concentration and not pathlength differences. This will actually introduce non-linearity's in the spectra-constituent correlation's and degrade the predictive ability of the model. However, if the concentration range is relatively small, this is a great way to build models that are insensitive to changes in pathlength of both the training spectra and spectra of "unknowns."

### Unit Area Normalization

This method attempts to correct the spectra for indeterminate pathlength when there is no way of measuring it, or isolating a band of a constant concentration constituent. In this approach, the spectra are normalized by calculating the area under the curve for the entire spectrum.

In effect, this method is the same as using the Thickness Correction on a large region of the spectrum as mentioned above. However, here the entire spectrum is always used, rather than a large selected region. This method is very simple to implement, but has some drawbacks. First, the concentration variations between all the training samples and "unknown" samples must not be too large, for the same reasons discussed in [Thickness Correction](#). In addition, since this method uses the entire spectrum, the

responses at all wavelengths in the spectrum must contain useful data. This means spectra that exhibit evidence of detector or optics cutoffs, or "black sample" (sample is too thick or too concentrated causing complete absorbance of all light at some wavelengths) cannot be corrected. Finally, if the spectra do not have a constant baseline between all measurements, the integrated area will be calculated incorrectly. It is generally best to combine this method with some form of baseline correction.

### **Correcting Baseline Effects**

As all spectroscopists know and have observed, spectrometers do not always collect data with an ideal baseline. Due to a variety of problems (detector drift, changing environmental conditions such as temperature, spectrometer purge, sampling accessories, etc.), the baseline of a given spectrum is not always where it should be. Beer's Law assumes that the absorption of light at a given wavelength is due entirely to the absorptivity of the constituents in the sample; it does not account for "spectrometer error" or "sampling error." Therefore, in order to accurately calculate concentrations, it is necessary to remove the baseline effect introduced by the spectrometer.

As with most random variations in the spectral data, most chemometric models can compensate for these effects by adding extra factors. Or, if the variations are truly completely random, ignore them altogether. However, as with all preprocessing methods, a more robust model will usually result when the known interference's in the data are removed first.

There are a number of methods used by spectroscopists to remove baseline effects from the spectra they collect. The problem with most methods is that they require the spectroscopist to decide that the baseline is correct by visual inspection. In addition to being very subjective, most of these methods cannot be easily applied in the somewhat automated fashion required for a calibration model.

However, there are some methods which are reasonably automated enough to be used as part of a calibration model. The following list of [baseline correction](#) methods is not exhaustive, and there are many other ways of auto-correcting the spectrum baseline as a chemometric preprocessing step.

### ***Linear Regression Baseline Fitting***

This is a very simple approach to baseline correction in that it requires no effort to set up. In this method, a least squares regression line is fit to the responses in each spectral region selected for calibration. This line is then subtracted from the response values in the region before using the data to perform the calibration model calculations.

Unfortunately, this is not always the best approach, especially when the selected spectral regions are primarily large bands from the constituents of interest. It tends to work better when the entire spectrum is used or when the selected regions are very broad. In some cases, this method actually degrades the performance to the calibration models more than if no baseline correction was used at all. In general, this method should only be used in situations where baseline aberrations are severe and a limited number of training sample spectra are available.

### ***Two Point Linear Baseline***

Another approach is the tried-and-true method of selecting two baseline points in the spectrum, connecting them with a line, and subtracting it from the spectral responses. This is known as a two-point baseline correction.

The main problem with this method is selecting the two points. In the optimum case, the training sample spectra will always have at least two regions that are at different ends of the spectrum where there is no absorption. In the worst case, the entire spectrum may exhibit absorption, and defining a baseline point can be difficult if not almost impossible. Another problem is how to select baseline

points from looking at a single spectrum, and be sure that no band will suddenly appear at that wavelength in future spectra.

Despite these limitations, there are some things that can be done to make this method of baseline correction more robust. For example, instead of selecting two single points for baseline correction, select a range of points in two parts of the spectrum. Then locate the wavelength point which has the minimum response in each range, and use these values as the two points. Another method is to calculate the average of the points in the selected baseline regions. These methods will presumably get around the problem of peaks shifting near the selected points.

### First and Second Derivatives

One of the best methods for removing baseline effects is the use of derivative spectra. This method is one of the earliest methods used to attempt to correct for baseline effects in spectra solely for the purpose of creating robust calibration models. The 1st derivative of a spectrum is simply a measure of the slope of the spectral curve at every point. The slope of the curve is not affected by baseline offsets in the spectrum, and thus the 1st derivative is a very effective method for removing baseline offsets. The 2nd derivative is a measure of the change in the slope of the curve. In addition to ignoring the offset, it is not affected by any linear "tilt" that may exist in the data, and is therefore a very effective method for removing both the baseline offset and slope from a spectrum.

There are many ways to calculate derivatives. One of the easiest is by using the method of simple differences. In this approach, the derivative at a given point is calculated by:

$$A_i = A_{i+1} - A_i$$

Unfortunately, this method is not always useful for calculating "real" derivatives. In fact, since it attempts to estimate the derivative from the between-point differences, in most cases it only succeeds in enhancing the noise in the spectrum.

There are better algorithms calculating derivatives including the [Gap method](#) and the [Savitzky-Golay method](#). Both of these algorithms use information from a localized segment of the spectrum to calculate the derivative at a particular wavelength rather than the difference between adjacent data points. In most cases, this avoids the problem of noise enhancement from the simple difference method and may actually apply some smoothing to the data.

One problem in applying these methods is that they require an extra parameter; the size of the spectral segment to use for calculation of the derivative points. For the Gap method, this is the size of the gap (usually measured in wavelength span, but sometimes in terms of data points) between the difference points. The Savitzky-Golay method uses a convolution function, and thus the number of data points in the function must be specified. If the segment is too small, the result may be no better than using the simple difference method. If it is too large, the derivative will not represent the local behavior of the spectrum (esp. Gap), and it will smooth out too much of the important information (esp. Savitzky-Golay). Although there have been many studies done on the appropriate size of the spectral segment to use, a good general rule is to use a sufficient number of points to cover the full width at half height of the largest absorbing band in the spectrum.

The main disadvantage of using derivative preprocessing is that the resulting spectra are very difficult to interpret. As mentioned above, the loading vectors for the calibration model represent the changes in the constituents of interest. In some cases (especially in the case of PLS-1 models), the vectors can be visually identified as representing a particular constituent. However, when derivative spectra are used, the loading vectors cannot be easily identified. In addition, the derivative makes visual interpretation of

the residual spectrum more difficult, and thus locating the spectral absorbencies of impurities in the samples cannot be done.

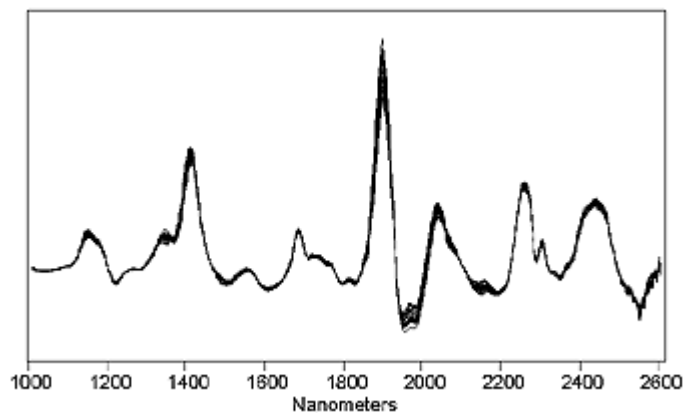


Figure 1. First derivatives of the spectra in Figure 3 below. Derivatives were calculated using the Gap method with a gap value of 10 nm.

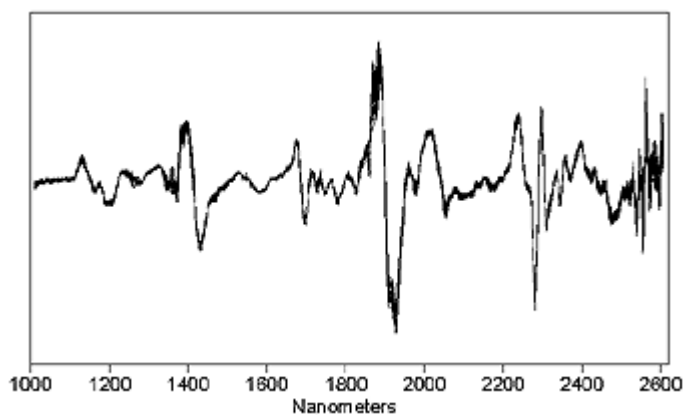


Figure 2. Second derivatives of the spectra in Figure 3 below. Derivatives were calculated using the Gap method with a gap value of 10 nm.

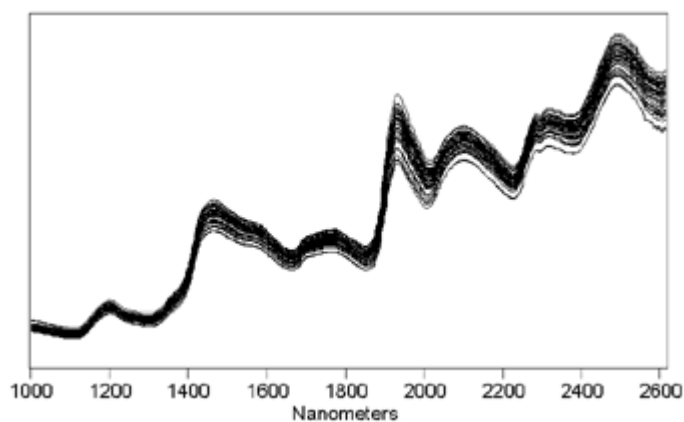


Figure3. A set of 50 [Log\(1/R\)](#) NIR spectra of ground wheat

samples measured using diffuse reflectance. The concentrations of the constituents of interest fall in a relatively narrow concentration range. However, note that the light scattering causes the spectra to appear quite different.

### Data Enhancement

Due to the multivariate nature of factor-based chemometric models, the direct relationship between the spectral response and the constituent concentration (univariate) is not very important. These models do not look at the absolute relationship between these values, but instead they calculate the relative change in the spectra and attempt to correlate that to a corresponding change in the constituent concentrations. This is why the models tend to be so robust and why they can calibrate for the constituents of interest in the presence of many other interference's.

Due to this fact, there are some mathematical enhancements that can be applied to data that is to be used in a multivariate model that would render it useless for a univariate model. The purpose of these algorithms is to remove redundant information and enhance the important sample-to-sample differences that exist within the data.

### Mean Centering

Mean centering is almost always applied when calculating any multivariate calibration model. This involves calculating the average spectrum of all the spectra in the training set and then subtracting the result from each spectrum. In addition, the mean concentration value for each constituent is calculated and subtracted from the concentrations of every sample.

$$\bar{A}_j = \sum_{i=1}^n A_{i,j}$$

Mean Spectrum:

$$\text{Mean Centering: } A_{i(MC)} = A_i - \bar{A}$$

In these equations,  $A$  is the  $n$  by  $p$  matrix of training set spectral responses for all the wavelengths,  $\bar{A}$  is a  $1$  by  $p$  vector of the average responses of all the training set spectra at each wavelength,  $A_j$  is a  $1$  by  $p$  vector of the responses for a single spectrum in the training set,  $n$  is the number of training spectra, and  $p$  is the number of wavelengths in the spectra.

By removing the mean from the data, the differences between the samples are substantially enhanced in terms of both concentration and spectral response. This usually leads to calibration models that give more accurate predictions.

### Variance Scaling

Variance scaling is used to emphasize small variations in the data by giving all values equal weighting. Variance scaling is calculated by dividing the response at each spectral data point by the standard deviation of the responses of all training spectra at that point. The concentration data is scaled likewise for each constituent. Note that variance scaling is only applicable after the data has already been mean centered.

$$Av_j = \sqrt{\frac{\sum_{i=1}^n (A_{i,j(MC)})^2}{(n-1)}}$$

Variance Spectrum:

$$\text{Variance Scaling: } A_{i(VS)} = A_{i(MC)} - Av$$

In these equations,  $A$  is the  $n$  by  $p$  matrix of training set spectral responses for all the wavelengths,  $Av$  is a  $1$  by  $p$  vector of the variance of the training set spectral responses at each wavelength,  $A_j$  is a  $1$  by  $p$  vector of the responses for a single spectrum in the training set,  $n$  is the number of training spectra, and  $p$  is the number of wavelengths in the spectra.

This preprocessing algorithm is most useful when analyzing minor (low concentration) constituents that have spectral bands that overlap those of major (higher concentration) constituents. By giving all the information in the data equal weighting, the calibration errors in the model should be more consistent across all constituents.

## Qualitative Spectroscopy Methods

[Discriminant Analysis](#)

[Discriminant Analysis, Mahalanobis Distance](#)

[Discriminant Analysis, The PCA/MDR Method](#)

[Spectral Library Search, Euclidean Distance Algorithm](#)

[Spectral Library Search, Absolute Value Algorithm](#)

[Spectral Library Search, Least Squares Algorithm](#)

[Spectral Library Search, First Derivative Absolute Value Algorithm](#)

[Spectral Library Search, First Derivative Least Squares Algorithm](#)

[Spectral Library Search, Correlation Algorithm](#)

[Spectral Library Search, First Derivative Correlation Algorithm](#)

[Spectral Library Search, Peak Matching](#)

## Quantitative Spectroscopy Methods

[Beer-Lambert Law](#)

[Beer-Lambert Law, Least Squares Regression](#)

[Beer-Lambert Law, Classical Least Squares \(K-Matrix\)](#)

[Beer-Lambert Law, Inverse Least Squares \(P-Matrix\)](#)

[Chemometrics, Preprocessing Techniques](#)

[Partial Least Squares \(PLS\)](#)

[Principal Component Analysis Methods](#)

[Principal Component Analysis Methods, Optimization](#)

[Calculating the Principal Components, The NIPALS Algorithm](#)

[Calculating the Principal Components, Decomposition of the Variance-Covariance Matrix](#)

[Principal Component Regression](#)

## References:

### Chemometrics, Preprocessing Techniques

P. Geladi, D. MacDougall and H. Martens, *Appl. Spec.*, 39, 491 (1985). (Reference to Multiplicative Scatter Correction.) K.H. Norris and P.C. Williams, *Cereal Chem.*, 62, 158 (1984). Karl H. Norris, *Food Research & Data Analysis, Proceedings of the 1982 IUFST Symposium*, H. Martens Ed., Applied Science Publishers, Oslo, 1983. (Using gap derivatives as pre-processing for quantitative models.) A. Savitzky and M.J.E. Golay, *Anal. Chem.*, 36, 1627 (1964). J. Steiner, Y Termonia and J. Deltour, *Anal. Chem.*, 44, 1906 (1972). H.M. Madden, *Anal. Chem.*, 50, 1383 (1978). (The original (top), the first correction (middle), and the second correction (bottom) for the Savitzky-Golay smoothing & derivative method.) B.J. Barnes M.S. Dhanoa and S.J. Lister, *Applied Spectroscopy*, 43, 772 (1989). (Using SNV & Detrending for quantitative models.) M.P. Fuller, G.L. Ritter and C.S. Draper, *Applied Spectroscopy*, 42, 217 (1988). (Reference to simple pathlength and auto-baseline correction preprocessing.)

### Classical Least Squares, (K-Matrix)

C.W. Brown, P.F. Lynch, R.J. Obremski and D.S. Lavery, *Analytical Chemistry*, 54, 1472, (1982). D.M. Haaland and R.G. Easterling, *Applied Spectroscopy*, 34, 539 (1980). D.M. Haaland and R.G. Easterling, *Applied Spectroscopy*, 36, 665 (1982). D.M. Haaland, R.G. Easterling and D.A. Vopicka, *Applied Spectroscopy*, 39, 73 (1985).

### Deconvolution

P. R. Griffiths and G. Pariente, *Trends in Analytical Chemistry* 5, No. 8, (1986).

### Derivative

K. H. Norris and P. C. Williams, *Cereal Chem*, 61 #2 (1984), 158. Madden, *Anal. Chem.*, 50 #9 (1978), 1383. Steiner et al., *Anal. Chem.*, 44 #11 (1972), 1906. Savitzky and Golay, *Analytical Chemistry* 36, 1627, (1964)

### Discriminant Analysis

H.L. Mark and D. Tunnell, *Analytical Chemistry*, 57, 1449, (1985). H.L. Mark, *Analytical Chemistry*, 58, 379, (1985). H.L. Mark, *Analytical Chemistry*, 59, 790, (1987). (references to using Mahalanobis distances in wavelength space.) P. C. Mahalanobis, *Proc. Natl. Institute of Science of India*, 2, 49, (1936) (Original reference to Mahalanobis distance calculations.) J.M. Dale and L.N. Klatt, *Appl. Spec.*, 43, 1399, (1989). M.F. Devaux, D. Bertrand, P. Robert and M. Qannari, *Appl. Spec.*, 42, 1015, (1988). G. Downey, P. Robert, D. Bertrand and P.M. Kelly, *Appl. Spec.*, 44, 150, (1990). J.S. Shenk, I. Landa, M.R. Hoover and M.O. Westerhaus, *Crop Sci.*, 21, 355, (1981). I.T. Joliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986. (References to other discrimination techniques as well as PCA.)

### Interpolation/Deresolution

W.A. Press, B.P. Flannery, et al, *Numerical Recipes*. Cambridge University Press. 1986. pg. 86-89.

### Interferogram Compute

Peter R. Griffiths and James A. de Haseth. "Fourier Transform Infrared Spectroscopy". John D. Wiley & Sons. 1986. Chapters 1 and 3. Robert H. Norton and Reinhard Beer. *Journal of the Optical Society of America*. Vol. 66, No. 3, March 1976. pg.259-64. Robert H. Norton and Reinhard Beer. *Journal of the Optical Society of America*. Vol. 67, No. 3, March 1977. pg.419, Errata Table.

### Inverse Least Squares, (Multiple Linear Regression, P-Matrix)

C.W. Brown, P.F.Lynch, R.J. Obremski and D.S. Lavery, *Analytical Chemistry*, 54, 1472, (1982). H. Mark, *Analytical Chemistry*, 58, 2814 (1986). K.H. Norris and P.C. Williams, *Cereal Chem.*, 62, 158 (1984).

### **PeakFitting by Non-linear Least Squares**

D.W. Marquardt, *J. Soc. Ind. Appl. Math.* **11**, pp. 431-441, (1963). Savitzky and Golay, *Analytical Chemistry* **36**, 1627, (1964). Steiner et al., *Analytical Chemistry* **44**, No. 11, pp.1906-1909 (1972). Madden, *Analytical Chemistry* **50**, No. 9, pp.1383-1386 (1978). P.D. Wilson and S.R. Polo, *J. Opt. Soc. Am.* **71**, No. 5, pp. 599-603 (1981). W.H.Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling in "Numerical Recipes"; Cambridge University Press, 1986; page 521-528. DE Metzler, CM Harris, et al., "Spectra of 3-Hydroxypyridines, Band-Shape Analysis and Evaluation of Tautomeric Equilibria"; *Biochemistry*, **12** #26 (1973) 5377.

### **Principal Component Analysis Methods**

E.H. Malinowski and D.G. Howery, *Factor Analysis in Chemistry*, John Wiley & Sons, New York, 1980. E.H. Malinowski, *J. Chemometrics*, 1, 33 (1987). E.H. Malinowski, *Analytical Chemistry*, 49, 606, (1977). (References to indicator functions, error analysis and factor selection methods for PCA.) I.T. Joliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986. R. Aries, D. Lidiard and R. Spragg, *Spectroscopy*, 5, 41 (1990). (A very nice tutorial on the basic principals behind factor analysis.) P.M. Fredericks, J.B. Lee, P.R. Osborn and D.A J. Swinkels, *Applied Spectroscopy*, 39, 303, (1985). S. Wold and M. Sjostrom, *Chemometrics: Theory and Application*, Kowalski, B.R., Ed., American Chemical Society, Washington, DC, 1977, pp 242-282. Y. Miyashita et al., *J. Chemometrics*, 4, 97 (1990). (Step-by-step description of the NIPALS PCA algorithm.)

### **Principal Component Regression/Partial Least Squares**

P.M. Fredericks, J.B. Lee, P.R. Osborn and D.A J. Swinkels, *Applied Spectroscopy*, 39, 303, (1985). (An excellent reference to PCR and PCA methods.) D.M. Haaland and E.V. Thomas, *Analytical Chemistry*, 60, 1193 (1988). D.M. Haaland and E.V. Thomas, *Analytical Chemistry*, 60, 1202 (1988). (Show worked out examples of the PLS algorithm. Also shows use of F-test for outlier detection.) P. Geladi and B. Kowalski, "Partial Least Squares Regression (PLS): a Tutorial", 1985 Educational Note, C.P.A.C. - University of Washington (1985). M. Martens and H. Martens, *Statistical Procedures in Food Research*, "Partial Least Squares Regression", J.R. Piggott Ed., Elsevier Appl. Sci. Publishing, London, 1986.

W. Lindberg, J. Persson and S. Wold, *Analytical Chemistry*, 55, 643 (1983).

(One of the first papers showing the full PLS algorithm. Caution; this paper has errors and omissions.) J. Sun, *J. Chemometrics*, 9, 21 (1995). T. Almøy, and E. Haugland, *Applied Spectroscopy*, 48, 327 (1994). J.M. Sutter, J.H. Kalivas and P.M. Lang, *J. Chemometrics*, 6, 217 (1992). (References to different methods of selecting individual vectors to use for PCR, a.k.a., Reduced PCR.) A. Lorber and B.R. Kowalski, *Applied Spectroscopy*, 44, 1464 (1990). (Reference to the Leverage validation method for PCR & PLS.)

### **Smoothing**

Steiner et al., *Analytical Chemistry*, **44** #11 (1972) 1906. Madden., *Analytical Chemistry*, **50** #9 (1978) 1383.

### **Spectral Library Search**

Alan Hanna, John C. Marshall and T. L. Isenhour, "A GC/FT-IR Compound Identification System", *J. Chrom. Sci.*, 17, 434 (1979). G. T. Rasmussen and T. L. Isenhour, "Library

Retrieval of Infrared Spectra Based on Detailed Intensity Information", *Applied Spectroscopy*, 33, 371, (1979). Stephen R. Lowry and David A. Huppler, "Infrared Spectral Search System for Gas Chromatography/Fourier Transform Infrared Spectrometry", *Analytical Chemistry*, 53, 889 (1981).

### **Spectral Subtraction**

P. C. Gillette and J. L. Koenig, *Applied Spectroscopy*, **38** #3 (1984), 334. M. A. Friese and S. Banerjee, *Applied Spectroscopy*, **46** #2 (1992), 246. S. Banerjee and K. Li, *Applied Spectroscopy*, **45** #6 (1991), 1047.

### **Spectral Unit Conversion, Kubelka-Munk**

P. Kubelka and F. Munk, *Z. Tech. Phys.*, 12, 593 (1931). P. Kubelka, *J. Opt. Soc. Am.*, 38, 448 (1948). M.P. Fuller and P.R. Griffith, *Analytical Chemistry*, 0, 1906 (1978).

### **Chemometric Links:**

The Analytical Chemistry Springboard - Chemometrics and Related Areas at <http://www.anachem.umu.se/cgi-bin/jumpstation.exe?Chemometrics>

Chemometrics from A to Z at <http://www.wiley.co.uk/wileychi/chemometrics/>  
Food Technology, Dairy and Food Science, Royal Vet. and Agricultural University (Denmark)

CPAC Home Page at <http://www.cpac.washington.edu/> for the Center for Process Analytical Chemistry, University of Washington

ChemoBro: A Searchable Database of Publications in Chemometrics at <http://www.optimax.dk/chemobro.html> the Food Technology, Dairy and Food Science, Royal Vet. and Agricultural University (Denmark)

North American Chapter of the International Chemometrics Society at <http://www.iac.tuwien.ac.at/NAmICS/WWW/welcome.html>

RG's Chemometrics Sites at <http://gepasi.dbs.aber.ac.uk/roy/sites/chemsite.htm>

Steve Brown's Chemometrics Review at <http://gopher.udel.edu/chemo/>

**ThermoGalactic one of the many thermo-companies has an excellent tutorial to chemometrics, and much of the data in this handout came from their website [www.galactic.com](http://www.galactic.com)**