TRYING TO UNDERSTAND CHEMOMETRICS

interpretation of spectra and modelling for analysis

Todays automatic equipment has brought a situation, where routinely large quantities of data are collected and have to be processed. On one side this causes a high degree of redundancy, where the important information first has to be extracted. On the other significant information first hast to be separated from unwanted date at high speeds. Hence Chemometrics!

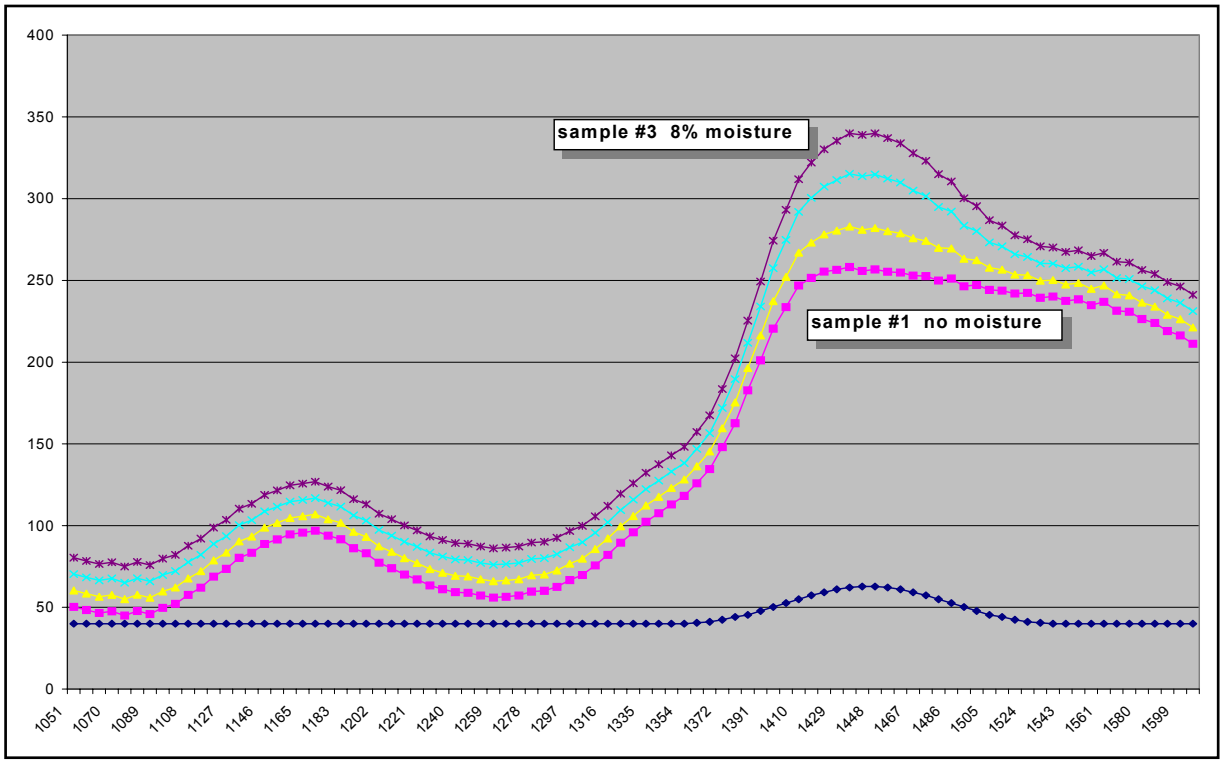
Traditionally understanding of spectroscopy either emerges from UV spectrocopy or from IR spectroscopy, where clearly separated and identifyable absorbance bands are found. The spectroscopist learns to think in terms of component associated absorbance peaks.

NIR spectroscopy however deals with weak and strongly overlapping absorbance bands of overtones or combinations of molecule vibrations, that cannot always be assigned. In addition they are influenced by many artifacts that lead to shifts and skews. Finally, variables thar are entered into an regression have to be linear independent from each other. Information has to be extracted from the set of spectra into an orthogonal (=independent) set of components or factors using a PCA (principal component analysis) or similar method. Although these factors look like spectra, it is important to understand that they are not spectra any more. The following is a small contribution to help understand this without mathmatics.

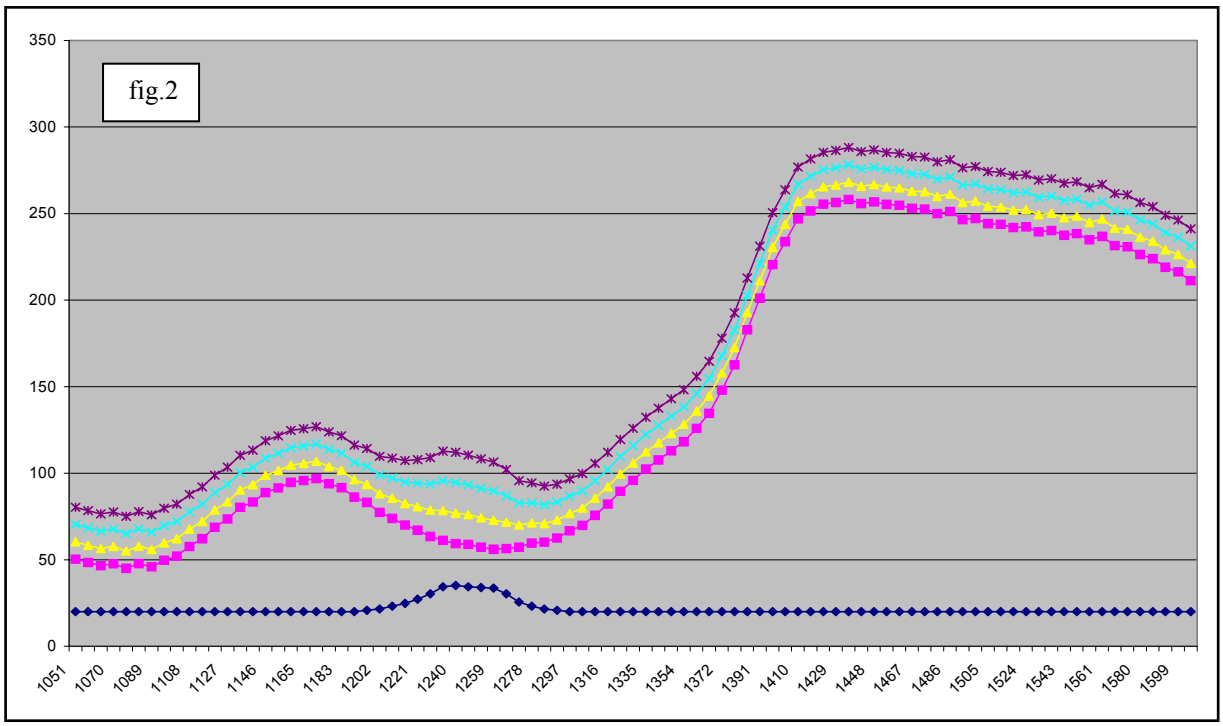
Principal Component Regression (just for understanding - very simplified !)

Let us assume, we have a set of spectra of our product with different content in moisture. one sample will contain no moisture, the others 3%, 6% and 8%. We know that moisture absorbs somewhere around 1450nm (first overtone) and therefore the spectra of these samples will look somewhat like this: (fig.1)

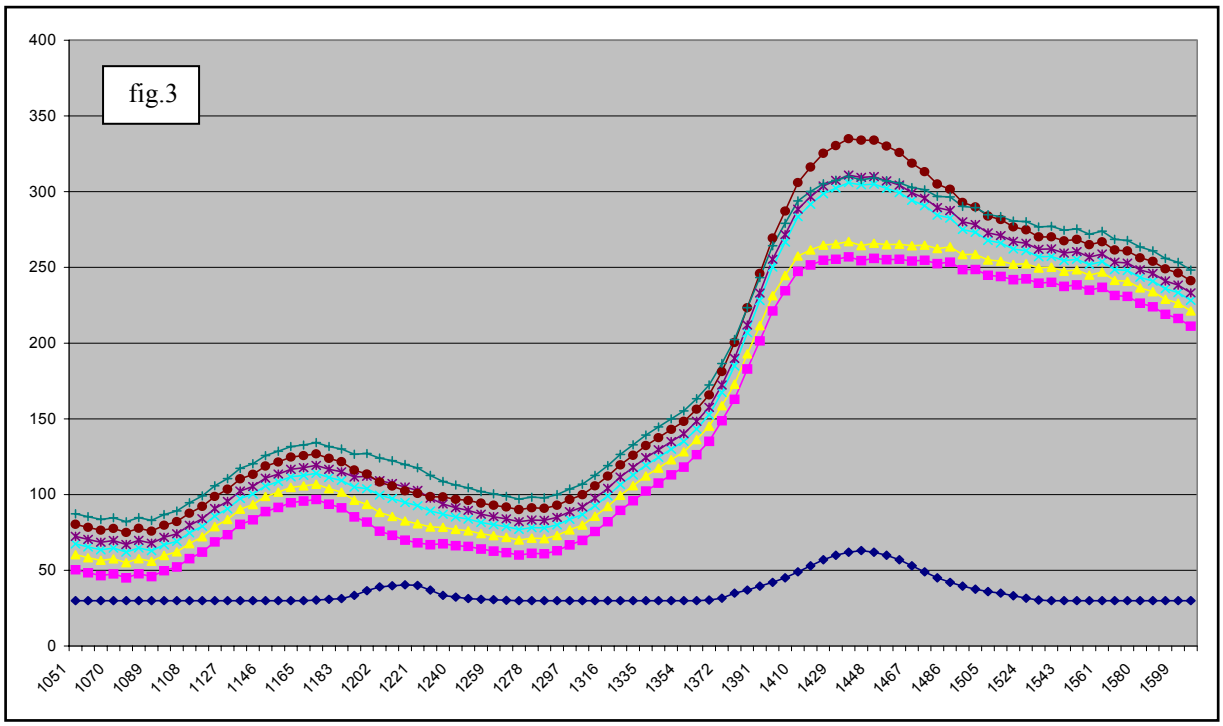
The bottom curve will be the graph indicating where in the spectra the differences will be found. (standard deviation of all current spectra)



The next set of spectra will be the same product with no water content but this time with different content of oil: 2%, 4% or 6%. This time the spectra (fig.2) are practically the same with some variations in the 1200nm range. Again, the bottom curve will be the graph indicating where in the spectra the differences will be found.

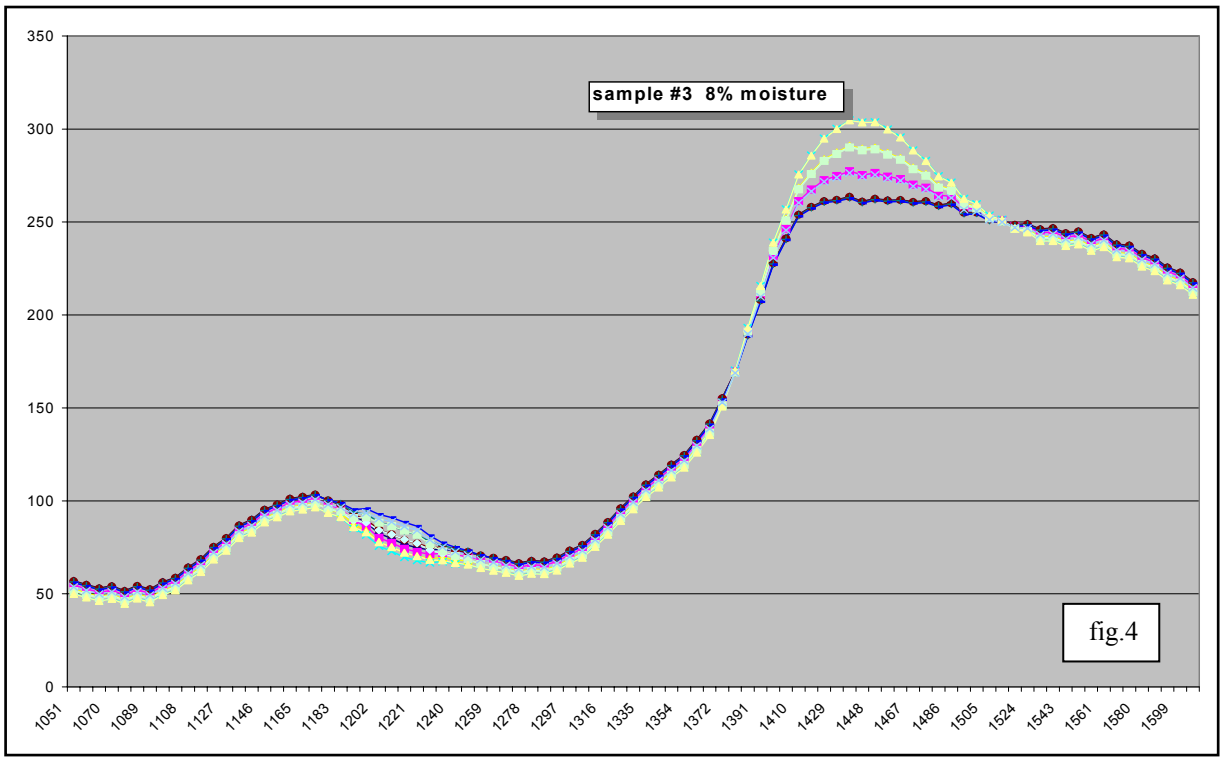


Now let's assume, we have spectra, that have variations in both the moisture and the oil content, again in aproxximately the same concentration range. Again, the bottom curve (fig.3) will be the graph indicating where in the spectra the differences will be found.

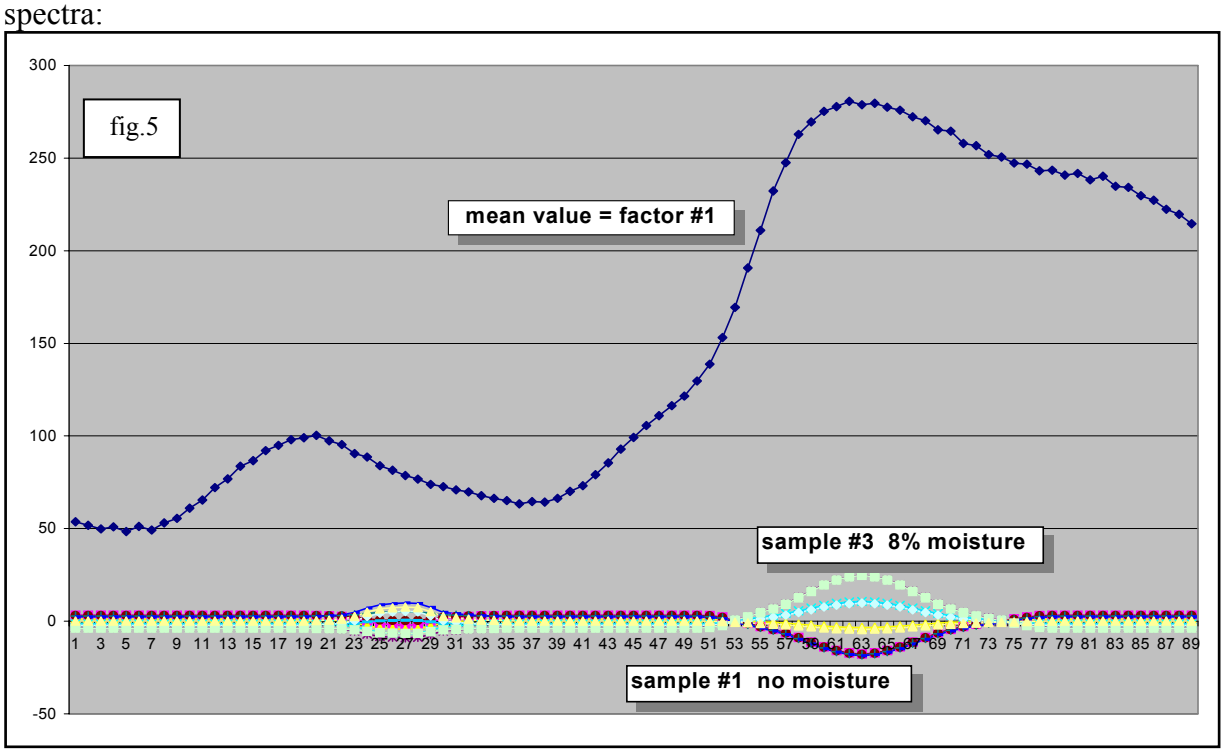


However it has to be noted that the two absorbances do not appear in all the spectra and not in the same intensities, as the sample contents of oil and moisture are independent from each other.

If now we do a mean centering as a spectra pre treatment and plot all the spectra we have in one graph (fig.4) we will have something to a real life data set of spectra:

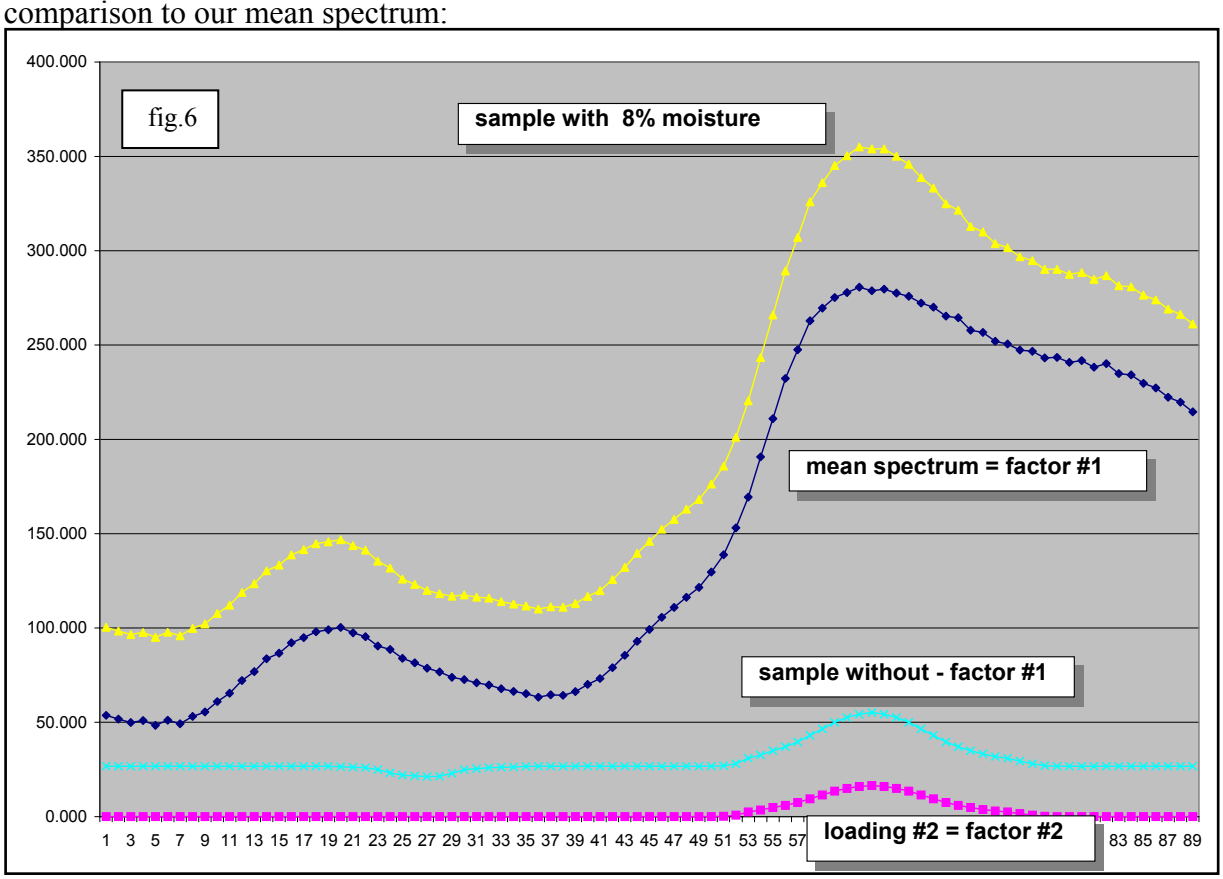


If we want to do a regression, which is the solving of a set of equations where on one hand we have our lab values and on the other the information from our spectra, we have to follow the rules for solving equations. An important one is, that equations have to be linear independent. However all the data points in the water absorption peak (see fig.2 – bottom curve) contain the same information and are not independent from each other and cannot be use as such for extracting the information.
We therefore have to extract the spectral information into independent factors. - But how?
If we look at the spectra, they all have a common shape. We could deduct the mean of all our spectra and call this the first factor. The remainder (fig.5) will contain all the variation in the spectra:



Now we extract the further information in such a way, that we find independent (= orthogonal) factors. The most important will be in the 1400nm region, with the highest (positive) peak for the 8% moisture sample and the lowest (most negative) for the 0% moisture. So we call this second factor that we extract a "loading", calculate the standard deviation and use it as a measure to calculate for each spectrum, how big the contribution (= "score") of this loading is. This is done for every spectrum, so each spectrum will have a "score value" attributed for for each "loading".

The following example shows this procedure Let us do this for the 8% moisture content in comparison to our mean spectrum:



First we take factor #1 (loading #1) and check, how much the influence of this factor is (= score of factor #1 in our spectrum). Assume the shape of our spectrum is close to the mean, so the score will be close to 1. The original spectrum now is reduced by 1 times factor #1 (= loading #1). We will now get our sample spectrum without the influence of the first factor and it will look something like the third graph. Now we take the loading #2 (the second factor, bottom trace) and check for best fit. Let's say we have to take 1.6 times this factor to get the best match, so we can say the score of the second factor in our spectrum is 1.6. The same will be done for the next factor #3 (which in our example will have a negative contribution as it goes below the zero line) and so on.

It now helps understanding if we realize, that we can regenerate every spectrum by taking each one of our factors (loadings), assigning the appropriate sacaling factor (score) and again adding all these components. The small remaining difference between our original spectrum and the regenerated one is the contribution of random noise (that cannot be modelled) and is called residuals. It is obvious that a spectrum that contains high moisture will have a higher score of the loading that contains most of the moisture information. We have found independant factors with no redundant information that can be used for a regression equation.

If now we do the calibration, we plot our scores against the reference lab values. As we would expect from the shape of our spectra and individual loadings, we will get the best explanation for our moisture content from loading number 2. If - as in our case – a spectrum with a score of 1.0 for this loading (the shape of the loading itself) means 6.5% moisture, our spectrum with about 8% has a score of 1.6, a spectrum with a score of 0.0 a moisture content of around 4%, one with a score of –0.5 a content of 2.7% and so on, then we can do a linear regression between the contribution (=score) of this factor (= loading) and the moisture content.

Our model now contains the shapes of our factors (= loadings) and a number, how much content (in our case water for loading #2) each score unit means. So if during prediction of results from a new spectrum we just have to determine how big the score of each loading is and know, what content this means.

Our model for prediction now consists of a set of loadings (usually three to six for PLS models and five to 10 for PCR models – depending on data pre treament procedures done). Multiplication with the new spectrum and adding will now yield the prediction value.

In real life it's not that easy as we do not get clearly assignable loadings and later factors somewhat correct the raw results we get from the first ones. Finally the last factors only will model random noise and have no more meaning. How many factors are important can only be found out by doing a validation (e.g. cross validation). As long as an additional loading contains valid information, it will improve the prediction quality. If an additional loading only models noise, the calibration still will improve, but when doing a prediction, the noise will deteriorate the quality of the result. This situation is called "overfitting". Therefore models can only be judged by doing prediction testing or at least cross validation.