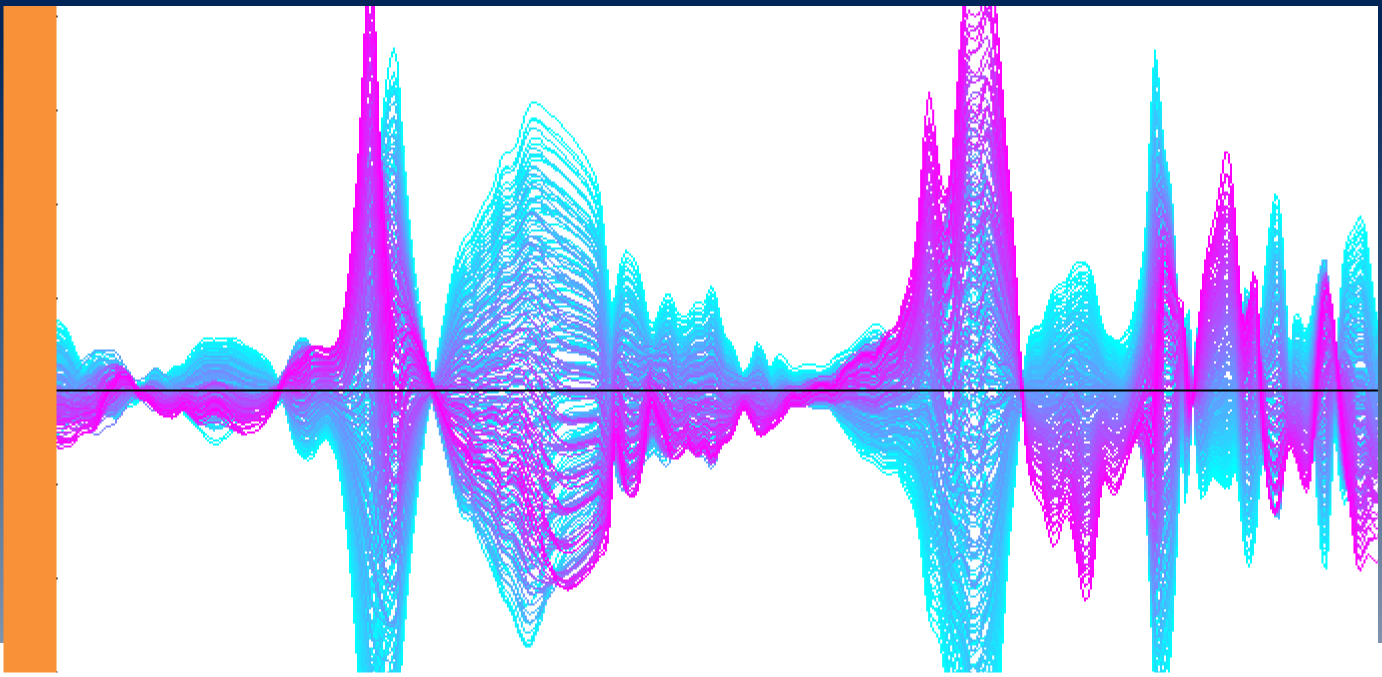

A white paper from FOSS

Principal Component Analysis and Near Infrared Spectroscopy

By: Lars Nørgaard, Senior Manager, Team Chemometric Development, FOSS
Rasmus Bro, Professor, University of Copenhagen, Denmark
Søren Balling Engelsen, Professor, University of Copenhagen, Denmark

Principal Component Analysis and Near Infrared Spectroscopy

In this new column in In Focus - the Chemometric Corner - different aspects of chemometrics will be addressed. We will start with a description of the work-horse of chemometrics – Principal Component Analysis and the application of the method on near infrared spectra.

Near infrared (NIR) spectroscopy has been the main driving force in the development of chemometrics since the seventies. NIR spectroscopy measures overtones and combination tones of the fundamental molecular vibrations in the infrared range, and especially, the asymmetric vibrations which are intensive in the near infrared range i.e. stretch vibrations involving hydrogen (e.g. C-H, O-H and N-H). These properties make NIR spectroscopy extremely useful for analysing all sorts of biological systems.

The fact that NIR spectroscopy measures the same basic molecular vibrations as a variety of overtones and combination tones of virtually the entire near infrared region, gives rise to strongly overlapping, almost holographic NIR spectra that are extremely difficult to interpret in the traditional manner.

NIR spectroscopic data are generally characterised by being highly co-linear, i.e. two adjacent wavelengths are normally positively correlated with high correlation coefficients. Principal Component Analysis (PCA) is a multivariate chemometric method that is optimal for handling co-linearity and as such, PCA and NIR spectroscopic data are a perfect match.

Mixture design

Application of PCA on NIR spectroscopic data is best illustrated with an example. For this purpose, we constructed a three-component mixture design with sucrose and its two monomer components, glucose and fructose, which were mixed with each component at 21 levels in the range 0% to 100% (Figure 1). Such a three component mixture design leads to a total of 21+20 +19 + ... + 1 = 231 mixtures, which were all measured with near infrared spectroscopy (NIR) in the range 1100 to 2500 nm.

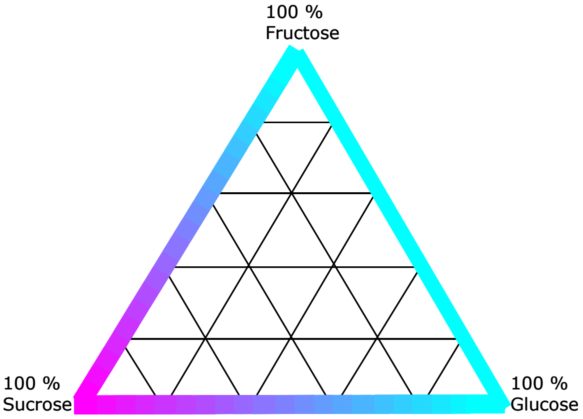

Figure 1. Mixture Design

Figure 2 shows the 231 NIR reflectance spectra recorded with a NIR spectrometer in reflectance mode against a white ceramic standard in the range 1100-2500 nm (FOSS NIRSystems 6500 placed at the University of Copenhagen). The unit on the ordinate is the log (1/R), where R is the ratio of the intensity reflected from the sample and the intensity reflected from the standard. The abscissa is the wavelength in nanometre (nm); every 4th nm is used corresponding to 350 spectral variables. Inspection of the NIR spectra shows that the spectra are not baseline separated, but consist of strongly overlapping peaks.

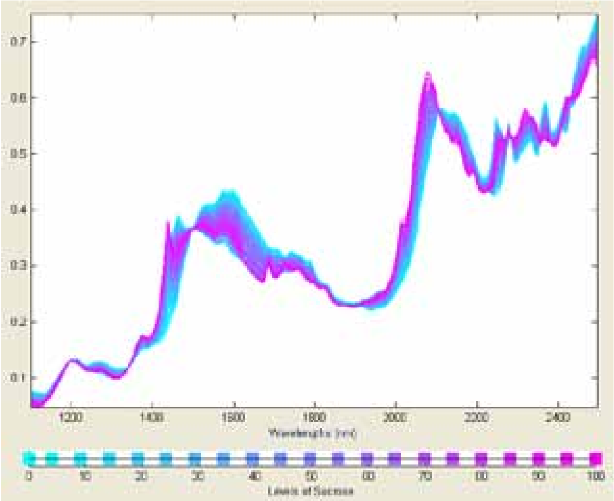

Figure 2. NIR spectra recorded on 231 mixture samples of sucrose, fructose and glucose. The spectra are coloured by the sucrose concentration (cyan is 0% and magenta is 100%).

PCA on NIR data

A PCA [1,2] model can be described as

Raw data = Mean Level + Model + Noise

Often, the first step in PCA modelling is to mean centre the spectroscopic data. This operation is performed to focus on the variations between individual samples rather than the absolute signal level (compare Figure 2 and 3). Mean centring is simply a subtraction of the average reflectance at each wavelength (Mean Level), so that the reflectance at each wavelength adds up to zero across all samples.

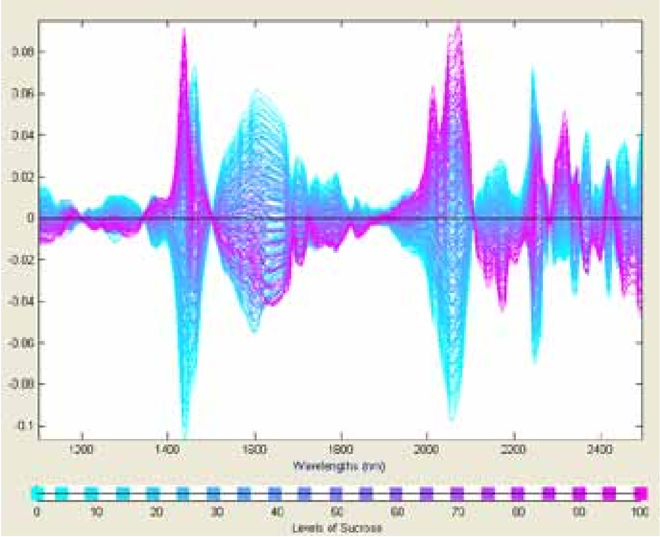

Figure 3 Mean centred NIR spectra, coloured according to the sucrose concentration.

In the example, the spectral data was pre-processed using the Multiplicative Signal Correction (MSC) [3] method, in order to correct for light scattering.

PCA - a linear and additive model

In Figure 4, the principle of PCA is illustrated for three selected samples. Please note that the PCA model is calculated on all 231 samples. To the left, in column one, the raw spectra are shown for sample 43 (blue), sample 107 (red) and sample 224 (green); the recorded spectra are presented as they are exported from the spectrometer software. Column two shows the average spectrum over all 231 samples; this is subtracted each individual sample spectrum, which is the mean centring part. The average spectrum is identical for all samples and therefore shown in the same colour for all samples (black**)**.

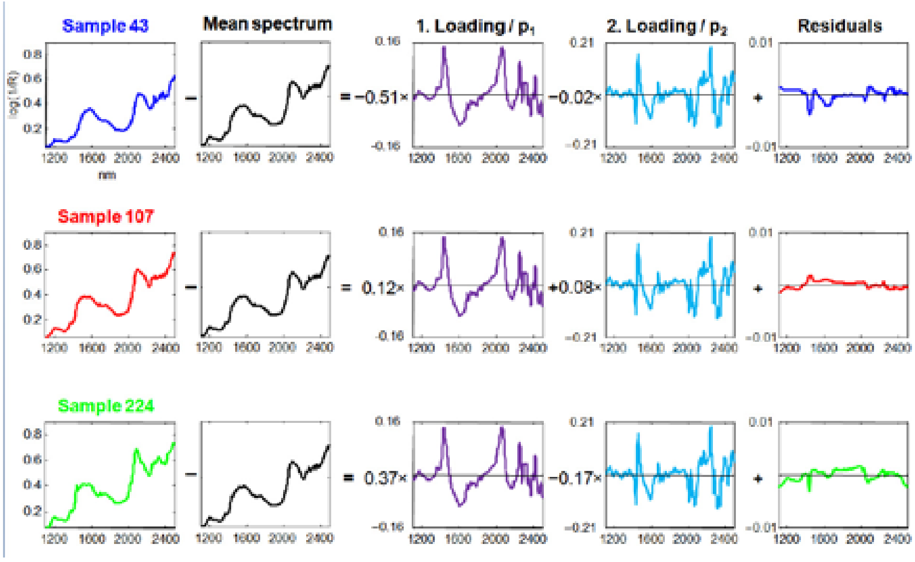

Figure 4. How PCA works. See text for a detailed description.

The first loading vector (purple) is the spectral structure that is best at describing the variation in the mean centred data (Figure 3). No other structure can explain more of the variation in the data than the first loading vector. The first loading is common to all samples; what makes the samples different is the content or concentration of this structure in their spectrum: this concentration is called the score value. For sample 43, the score value is -0.51 for the first loading. The 230 remaining samples in the data set have different score values. Multiplying the loading vector with -0.51 is the best description one can obtain for sample 43, when the loading vector should also describe the other samples.

The second loading (light blue) is the structure that describes the second most variation in the data set. The vector has the special property of being orthogonal (perpendicular) to the first loading. Once again, the sample diversity is reflected in the score value, which is -0.02 for sample 43.

Residuals and variance explained

The part of the variation in the data set not described by the first two loading vectors is represented in the residuals (Figure 4, column five). The residuals are specific for each sample and can be used for the detection of deviating sample patterns. Note that the numerical residual values vary within + / - 0.002 (the ordinate). These values can be compared directly with the original variation in the mean centred spectral data (Figure 3), which varies between -0.1 and +0.09.

By comparing the size of the residuals with the variation of the mean centred data, one can calculate the variance explained for each principal component. In this case the first component explains 88.0% of the total variation; the second component 11.6% of the variation, and the sum of the two components explains 99.6% of the total variation in the data set.

Scores plot

In this example we inspect only the first two principal components, which makes sense in relation to the number of chemical sources of variation in the samples: three chemical components in a mixture design (sum is 100%) ideally gives rise to two independent sources of variation.

By plotting all 231 scores for the first principal component versus the corresponding values for the second component we obtain a score plot (Figure 5). Each point in this score plot represents an NIR spectrum based on the original 350 measured variables. Sample 43 is placed in the coordinate system with coordinates (-0.51, -0.02); sample 107 at (0.12, 0.08) and so forth for the remaining 229 samples.

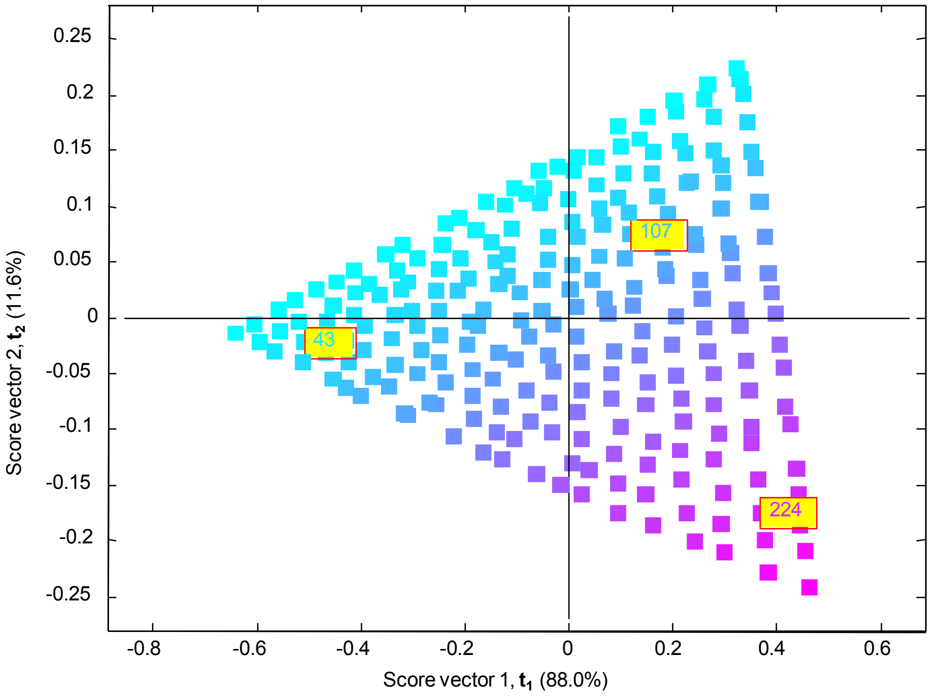

Figure 5. Scores plot from a PCA model on the NIR data. The mixture design pattern is evident. The samples are coloured by the sucrose concentration (cyan is 0% and magenta is 100%).

The mathematical model

A more precise mathematical model formulation of PCA is

$$\mathbf{X} = \mathbf{1x}'_{\text{average}} + \mathrm{T}_a\mathbf{P}'_a + \mathbf{E}_a$$

where $\mathbf{X}$ is the raw data, $\mathbf{1}$ is a column vector consisting of ones, $\mathbf{x}'_{\text{average}}$ is a row vector, which is the average of all the samples (= mean spectrum); $\mathbf{T}_a$ and $\mathbf{P}_a$ contain scores and loadings, and $\mathbf{E}_a$ is the residuals, i.e. the part of the original data not described by the model. Index a indicates the number of principal components calculated in the model, and ' indicates a transposed matrix.
The mean spectrum can be moved to the left side of the equation above, thereby making it possible to obtain the following description of the PCA model integrating the mean centring step into the raw data

$$\mathbf{X}_{\text{mean centred}} = \mathbf{X} - \mathbf{1x}'_{\text{average}} = \mathbf{T}_a\mathbf{P}'_a + \mathbf{E}_a$$

where $\mathbf{X}_{\text{mean centred}}$ are the mean centred spectral data with the same dimensions as the original $\mathbf{X}$ matrix, i.e. a data table with 231 samples and 350 variables, and the content of $\mathbf{T}_a$, $\mathbf{P}_a$ and $\mathbf{E}_a$ is given above.

For a two-component model the equation can be written as

$$\mathbf{X}_{\text{mean centred}} = \mathbf{T}_a\mathbf{P}'_a + \mathbf{E}_a = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_1' + \mathbf{E}_a$$

which is exactly the model that is illustrated in Figure 4.

In order to calculate a PCA model, several different algorithmic approaches can be applied. Several of these are implemented in commercial software packages that offer the possibility to both calculate and present results from a PCA model.

PCA and Lambert-Beer's law

PCA is superior for handling the highly co-linear data often found in spectroscopy. As the example illustrates, PCA is an excellent tool for exploratory data analysis: one can investigate the behaviour and characteristics of individual samples, and study the wavelength regions that are important for the similarity/difference between samples.

PCA can be thought of as a 'reverse' Lambert-Beer law, which is based on linearity and additivity: from measurements on whole spectra consisting of many contributions from chemical components in the sample, the method estimates the latent spectra (loadings) and determines the corresponding concentrations in the samples (scores) from the measured spectra.

Acknowledgements

The authors are grateful to Hanne Winning for compiling the data presented in the column.

References

[1] Hotelling H. Analysis of a Complex of Statistical Variables with Principal Components. Journal of Educational Psychology 24: 417-441 & 498-520, 1933.

[2] Wold S., Esbensen K., Geladi P. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 2: 37-52, 1987.

[3] Geladi P., MacDougall D., Martens H. Linearization and scatter-correction for near-infrared reflectance spectra of meat.
Applied Spectroscopy 39(3): 491-500, 1985.