Bringing the Chemical Knowledge into Empirical Models

Euroanalysis September 6 -10 2009, Innsbruck, Austria

Frank Westad: fw@camo.no | CAMO Software

INTRODUCTION

It is often desired to measure a number of samples from a set of experiments with several instrumental methods, e.g. in metabolomics. One of the objectives in such situations is to find out what is the chemistry that is described commonly by all instruments, and what are the contributions from individual instruments. At the same time, one wants to apply the chemical background knowledge to confirm the interpretation from the models.

The chemical background knowledge can be presented as qualitative information as to where in the spectrum (specific wave number regions) various functional groups such as O-H, C-H, N-H give signals (vibrational bands). Although these data tables have no common dimension as they are in a so-called L-shape they can be modelled with L-PLS regression1 This enables a visualization of the underlying chemistry to confirm the findings in the empirical models. As Near-infrared spectroscopy (NIR) reflects overtones and combination bands of Infrared spectroscopy (IR), it is a good example for demonstrating how the L-PLS regression method works.

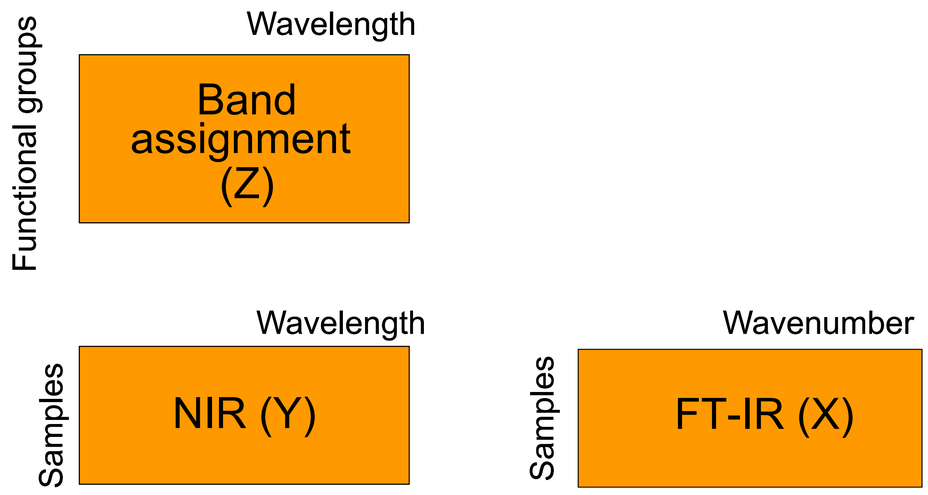

Figure 1. Illustration of the L-shaped data structure

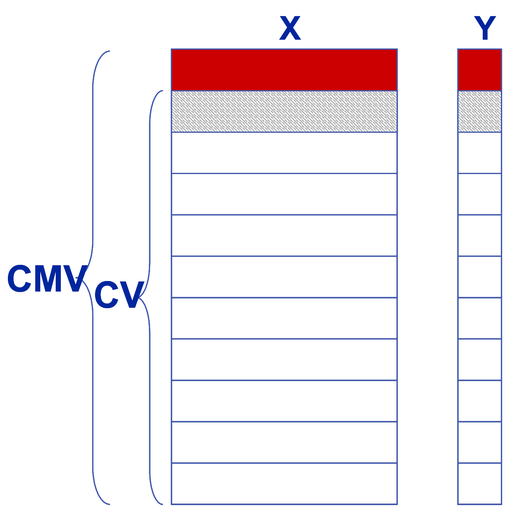

Figure 2. Illustration of cross model validation (CMV).

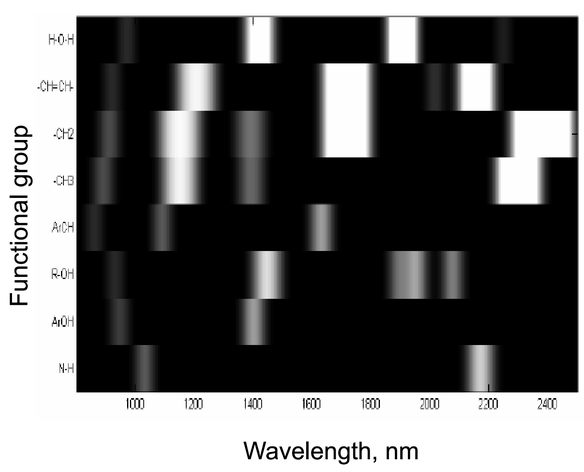

Figure 3. Band assignment for NIR spectroscopy

Methods:

PLS regression, jack-knifing and validation

Let the Partial Least Squares Regression (PLSR) be represented in matrix notation as:

$\mathbf{X} = \mathbf{TP}_T + \mathbf{E}_A$; $\mathbf{Y} = \mathbf{TQ}^T + \mathbf{F}_A$; $\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T$ in the regression equation $\overset{A}{\mathbf{Y}} = \mathbf{XB}$

Cross model validation with jack-knife estimates of B is a conservative approach for finding significant variables2

L-PLS regression

The simultaneous modelling of **X, Y and Z** can be achieved by singular value decomposition of the product $\mathbf{X}^T\mathbf{YZ}^T$. From this, models for **X, Y and Z** can be estimated and presented similar to PLSR.

$\mathbf{X} = \mathbf{T}_X\mathbf{P}_X\mathbf{T} + \mathbf{E}_X$, $\mathbf{Z} = \mathbf{T}_Z\mathbf{P}_Z\mathbf{T} + \mathbf{E}_Z$, $\mathbf{Y} = \mathbf{T}_X\mathbf{CT}_Z^T + \mathbf{F}$

A band assignment matrix for Near/Infrared spectroscopy was constructed by looking up known vibrational bands in the literature and represent these as qualitative information in the matrix **Z**. The matrix was convolved with a gaussian filter to reflect peak intensities.

Data:

- 32 samples of marzipan
- FT-IR (1700-600 cm-1) and VIS/NIR (400-2500 nm) spectroscopy.
- Preprocessing the VIS/NIR spectra with signal normal variate (SNV; i.e. center and scale objects to unit variance)
- Sugar content ranged from 35-70%
- The water content varied from 6-18%

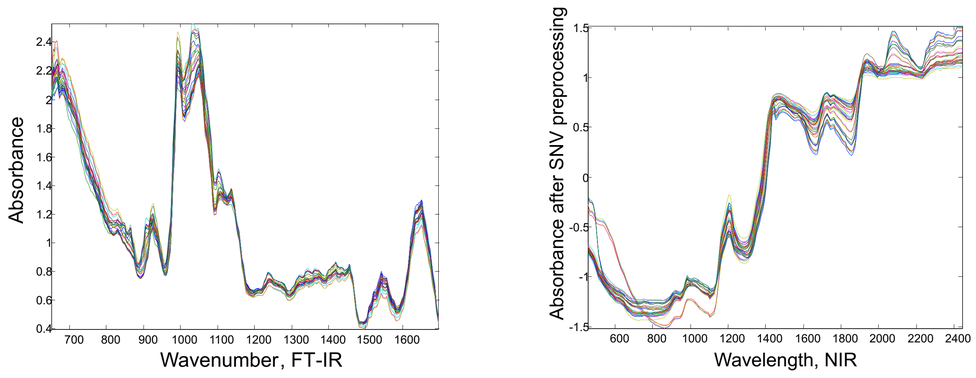

Figure 3. FT- IR spectra (left) and VIS/NIR spectra (right) of marzipan sample

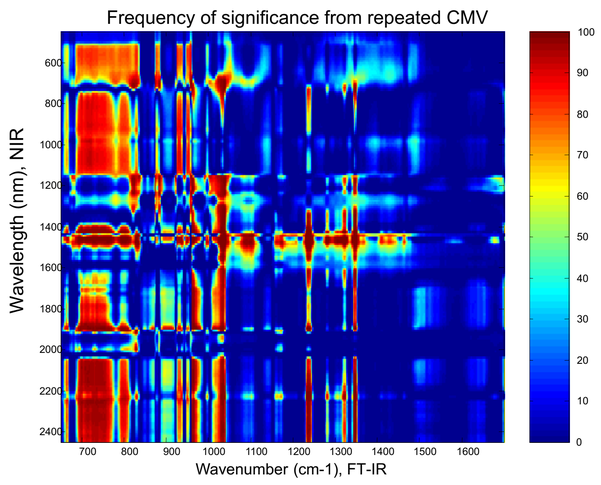

Figure 4. Map of significant regions between FT-IR and NIR spectra

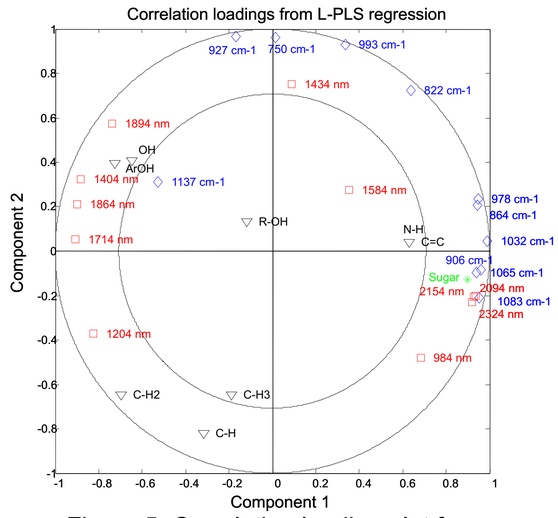

Figure 5. Correlation loading plot from L-PLS regression for selected bands

Results:

The 32 marzipan samples were subject to repeated CMV (100 runs. Uncertainty estimates from jack-knifing). 3 PLSR components was the basis for significance tests at 5% level. The main results are shown as a map of frequency of significance for the regression coefficient matrix B. Sugar and water constitute the chemical compounds of interest. Fig. 4 reveals a distinction between O-H vibrations (sugar and water, 900, 1450, 1940 nm) and C-H vibrations (sugar). C-H deformation at 820 cm- 1 correspond to the NIR region around 2200 nm. The region 1400–1500 nm is related to both O-H stretch vibrations and C-H combinations.

The main peaks in the bands that were significant were selected as input to the L-PLS regression. Figure 5 shows how correlation loadings enables a direct interpretation of the chemistry and the actual spectral regions that were found to be significant. Dummy variables for the samples can also be included (not shown), as can the sugar content, since correlation is independent of the scaling of variables. Note that sugar and water content are inversely correlated in the samples themselves, so there is both a concentration dependence as % content as well as a chemical dependence in terms of OH bands.

Conclusions:

- Cross model validation with jack-knife estimates is an efficient way of removing variables that are not of interest, and the relation between two instrumental methods can be presented as a color image
- L-PLS regression gives direct interpretation of the underlying chemistry which is useful for confirming existing knowledge but also for finding unknown phenomena which may lead to innovation and further research.
- The correlation loading plot is a very condensed way of visualizing all of interest for the three data tables as well as constituents such as water, sugar, fat.
- The method is general and can be applied for genetics (microarray, SNP) spectroscopy (FT-IR, NIR, Raman, NMR), and other types of data where the variables have known characteristics from the basic theory, e.g. chemistry and biology.

References:

- Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L.H., Høy, M., Westad, F., Thybo, A. & Martens, M. Regressing a matrix on descriptors of both its rows and of its columns, by low-rank L-PLS Regression. Computational Statistics and Data Analysis, 48, 103-125, 2005.
- Anderssen, E., Dyrstad, K., Westad, F. & Martens, M. Reducing over-optimism in variable selection by cross model validation. Chemometrics and Intelligent Laboratory Systems, 84 (1-2 SPEC. ISS.), pp. 69-74, 2006.
- Westad, F., Afseth, N.K., Bro, R. Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression. Analytica Chimica Acta, 595 (1-2 SPEC. ISS.), pp. 323-327, 2007.