

O2PLS[®] for improved analysis and visualization of complex data

Lennart Eriksson¹, Svante Wold² and Johan Trygg³

¹ Umetrics AB, POB 7960, SE-907 19 Umeå, Sweden, lennart.eriksson@umetrics.com

² Umetrics Inc., 17 Kiel Ave., Kinnelon, NJ 07405, USA, svante.wold@umetrics.com

³ Institute of Chemistry, Umeå University, SE-901 87 Umeå, Sweden, johan.trygg@chem.umu.se

Keywords: O2PLS, predictive components, Y-orthogonal variability, X-orthogonal variability.

1 Introduction

O2PLS is a generalization of PLS and OPLS [1-5]. In contrast to PLS and OPLS, it is bidirectional, i.e. $X \leftrightarrow Y$; therefore X can be used to predict Y, and Y can be used to predict X. O2PLS allows the partitioning of the systematic variability in X and Y into three parts: the X/Y joint predictive variation; the Y-orthogonal variation in X; and the X-unrelated variation in Y (Figure 1).

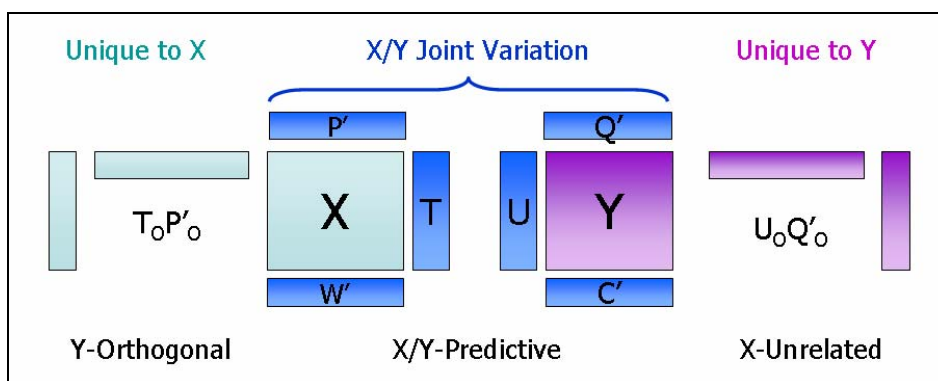


Figure 1. Overview of the O2PLS model relating two data tables to each other.

2 Theory

The O2PLS model can be written as:

$$\text{Model of X:} \quad X = T_p P'_p + T_o P'_o + E \quad (1)$$

$$\text{Model of Y:} \quad Y = U_p Q'_p + U_o Q'_o + F \quad (2)$$

where a linear relationship exists between T_p and U_p and the score vectors in T_p and T_o are mutually orthogonal. The number of components in the respective set of components is determined using cross-validation.

3 Material and methods

The application data set contains quantitative data for five different types of carrageenans derived from their NIR, IR and Raman spectra. Carrageenans are polysaccharides that are extracted from seaweed and used as gelling and thickening agents in a wide range of industries, including food, pharmaceuticals and cosmetics. Many different types of carrageenans exist, each having different gelling and thickening properties. The raw material (seaweed) contains a mixture of carrageenan types and hence the final commercial product is also a mixture of types. It is imperative that the industry know the composition of specific products in order to target appropriate applications areas or, if necessary, perform chemical modification prior to release.

The proportions of five different types of carrageenans (Lambda, Kappa, Iota, Mu and Nu carrageenans) were varied in this work using a five-component mixture design in six levels [6]. This produced a data set containing 128 samples [6],

sampled over five days. For each sample, NIR (1100 – 2500 nm; 699 variables), IR (550 – 4000 cm⁻¹; 662 variables) and Raman (3600 – 200 cm⁻¹; 3401 variables) spectra were acquired. These spectra are treated as three separate blocks of data. The relationship between these blocks and the proportions of the five carrageenan types in each mixture sample, is reported elsewhere [7].

The objective of this work is to investigate the information overlap between the three blocks of spectral data using O2PLS. Additionally, these analyses will reveal spectral variability unique to each method.

4 Results and discussion

The blocks NIR, IR and Raman data were analyzed in a pair-wise fashion using O2PLS as implemented in SIMCA-P⁺ version 12. NIR data (used as the X-block) and the IR data (used as the Y-block) were contrasted, and the results are given below. The O2PLS model obtained was a 6 + 1 + 3 model (Figure 2). The notation should be viewed as 6 predictive components taking care of the joint NIR/IR variation, 1 Y-orthogonal component expressing the variability in the NIR data that is not present in the IR data, and 3 X-unrelated components representing the variability in the IR data that is not available in the NIR data.

A		R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Q2(cum)	Significance
Σ	Model		0.997			0.713		0.671	
0	Cent.				Cent.				
P 1		0.48	0.494	63.2	0.286	0.287	0.219	0.219	R1
P 2		0.298	0.792	38.2	0.152	0.439	0.115	0.335	R1
P 3		0.13	0.922	16.6	0.166	0.605	0.141	0.476	R1
P 4		0.0597	0.982	7.64	0.0867	0.692	0.114	0.59	R1
P 5		0.0139	0.996	1.78	0.00769	0.7	0.0389	0.629	R1
P 6		0.00107	0.997	0.136	0.0136	0.713	0.042	0.671	R1
Σ	Orthogonal		0.0138			0.00046			
O 1		0.0138	0.0138	1.76	0.00046	0.00046			R1
Σ	Unrelated to X					0.178			

Figure 2. Model summary of the O2PLS model relating the NIR data to the IR data.

The results in Figure 2 show a high degree of information overlap between the NIR and IR data. Only 1.4% of the variability in the NIR data is orthogonal (unrelated) to the IR data while the analogous fraction in the IR data that is unrelated to the NIR data is larger, i.e. 17.8%.

Interpretation of the joint X/Y co-variation (the information overlap between the NIR and IR data):

The score plot (Figure 3) shows the data structure captured by the first two predictive components. The triangular distribution of the 128 samples is due to the underlying mixture design. This distribution indicates that the information overlap between the NIR and the IR data is affected by the systematically changing nature of the samples.

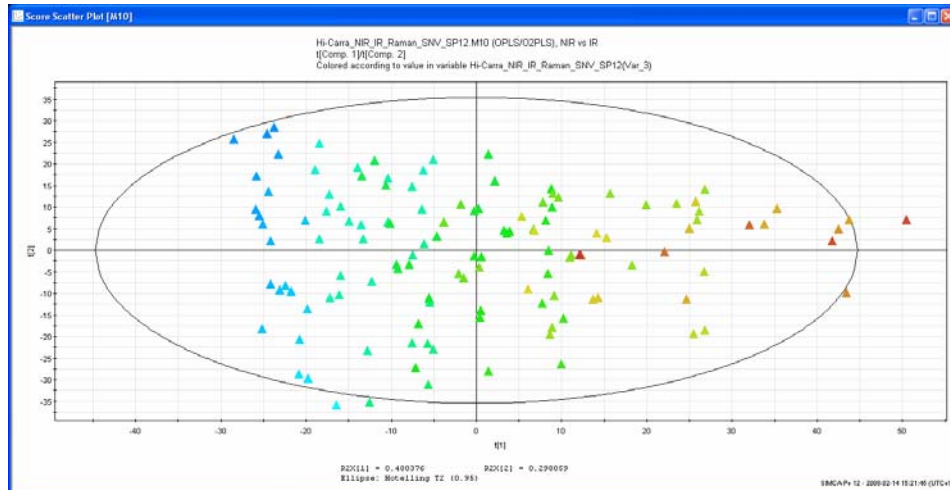


Figure 3. Scatter plot of the scores of the first two predictive components of the O2PLS model. Each point is one sample. The samples are colored according to the content of the Iota-type carrageenan constituent.

Interpretation of the Y-orthogonal variation in X (variation unique to the NIR data):

The variability in the NIR data that is orthogonal to the IR data amounts to 1.4%. An examination of the single score vector belonging to this compartment of the O2PLS model shows no apparent trend or grouping in the data. This suggests that the Y-orthogonal variation is spread in a similar fashion over the five sampling days. Figure 4 shows the loading spectrum of this component. The non-correlating variability mainly resides in the wavelength regions 1700-1860nm, 1950nm, and 2220-2350nm, whereas the region between 1100-1700nm contains comparatively little variability unique to the NIR data.

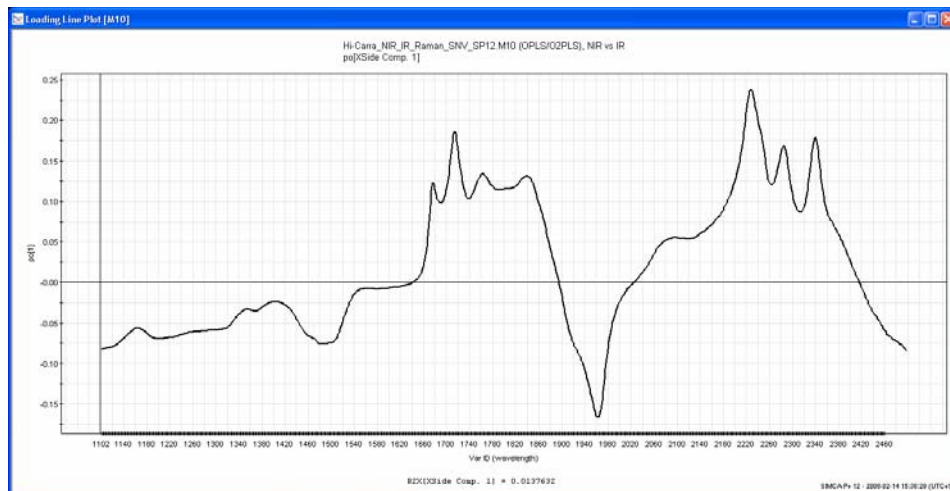


Figure 4. Loading spectrum of the single Y-orthogonal component. This plot highlights which spectral regions in the NIR data contain variability that is not present in the IR data.

Interpretation of the X-unrelated variability in Y (variation unique to the IR data):

The fraction of variability unique to the IR data is 17.8%. It is expressed by the three X-unrelated O2PLS components. A plot of the scores of the first of these components (Figure 5) shows a systematic shift among the data measured early in the sampling. This variability corresponds to 9.4% of the total variance explained and hence cannot be neglected. Such systematic differences are not seen in the NIR data for the same samples. The corresponding loading spectrum points to the wavenumber regions that capture this structure (Figure 6).

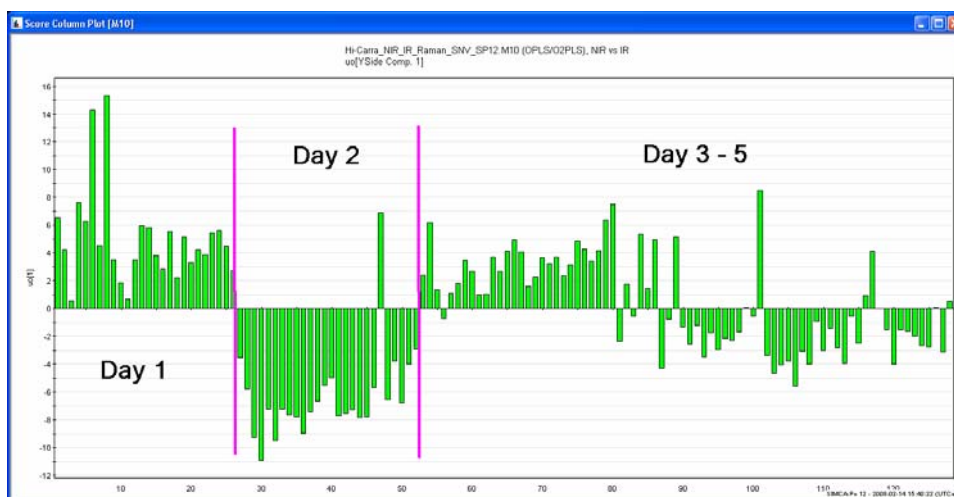


Figure 5. Score plot of the first X-unrelated O2PLS component. The horizontal lines indicate the shift in sample properties that takes place when going from day 1 to day 2, and from day 2 to 3.

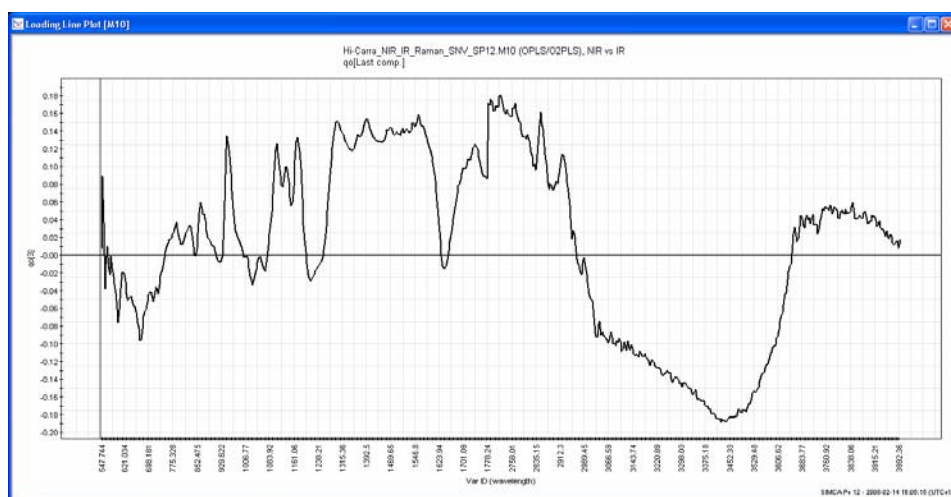


Figure 6. Loading spectrum of the first X-unrelated component. This plot highlights which spectral regions in the IR data contain variability that is not present in the NIR-data, i.e., where the deviating properties of the day 2 samples are seen.

5 Conclusion

The main advantage of the O2PLS method is that it simplifies the analysis, visualization and interpretation of complex, multi-block data sets by producing more informative plots than the conventional PLS method. This makes O2PLS highly useful for the treatment of analytical and bioanalytical chemical data, e.g. for comparing and contrasting blocks of data compiled using different spectral methods or different 'omics' platforms (microarray data, electrophoresis data, etc.). O2PLS is especially useful in calibration transfer applications where the method helps to expose systematic differences between analytical instruments. O2PLS is also useful in differentiating subject responses before and after a given treatment.

6 References

- [1] Trygg, J., and Wold, S., Orthogonal Projections to Latent Structures (OPLS), *Journal of Chemometrics*, 16, 119-128, 2002.
- [2] Trygg, J., Prediction and Spectral Profile Estimation in Multivariate Calibration, *Journal of Chemometrics*, 18, 166-172, 2004.
- [3] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, M., and Wold, S., Multi- and Megavariate Data Analysis, Part II, Method Extensions and Advanced Applications, Chapter 23, Umetrics Academy, 2005.
- [4] Trygg, J., O2-PLS for Qualitative and Quantitative Analysis in Multivariate Calibration, *Journal of Chemometrics*, 16, 283-293, 2002.
- [5] Trygg, J., and Wold, S., O2-PLS, a Two-Block (X-Y) Latent Variable Regression (LVR) Method With an Integral OSC Filter, *Journal of Chemometrics*, 17, 53-64, 2003.
- [6] Dyrby, M., Petersen, R.V., Larsen, J., Rudolf, B., Nørgaard, L., Engelsen, S.B., Towards on-line monitoring of the composition of commercial Carrageenan powders, *Carbohydrate Polymers*, 57, 337-348, 2004.
- [7] Eriksson, L., Dyrby, M., Trygg, J., and Wold, S., Separating Y-predictive and Y-orthogonal variation in multi-block spectral data, *Journal of Chemometrics*, 20, 352-361, 2006.