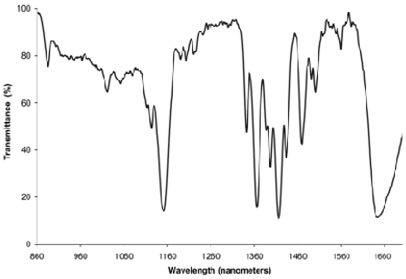
Multivariate modeling of spectroscopic data

NIR spectroscopy



Aim

- To find a correlation between spectroscopic information and some physico-chemical properties of the system
- Simple linear regression can not be used
- Some multivariate procedure should be used to address this problem.

Multivariate linear regression (MLR)

$$\left.\begin{array}{l} y_1 = b_0 + b_1u_{11} + b_2u_{12} + ... + b_pu_{1p} \\ y_2 = b_0 + b_1u_{21} + b_2u_{22} + ... + b_pu_{2p} \\ ... \\ y_n = b_0 + b_1u_{n1} + b_2u_{n2} + ... + b_pu_{np} \end{array}\right\}$$

Matrix form

$$|y_1 \quad y_2 \quad ... \quad y_n| = \left\| \begin{array}{ccccc} 1 & u_{11} & u_{12} & ... & u_{1p} \\ 1 & u_{21} & u_{22} & ... & u_{2p} \\ ... & ... & ... & ... & ... \\ 1 & u_{n1} & u_{n2} & ... & u_{np} \end{array} \right\| \left| \begin{array}{c} b_0 \\ b_1 \\ ... \\ b_p \end{array} \right|$$

$$\overset{\omega}{Y}^T = \| U \| \overset{\omega}{B} \Leftarrow$$

Vector form.

The number of examples must be equal or greater than the number of coefficients

Problems of MLR procedure

- The number of measurements should be greater or equal to the number of descriptors
- The colinearity of the descriptors.

Solution

- The reduction of the number of the descriptors should be performed

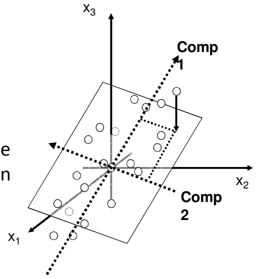
Principal Components Analysis - PCA

Tasks

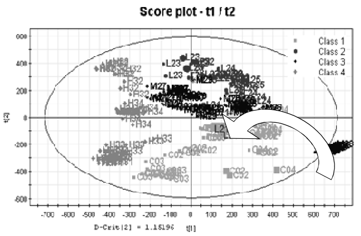
- Multivariate projection technique
- Dimensionality reduction
- Graphical overview

Advantages

- Plot data in K-Dimensional space
- Directions of maximum variation
- Orthogonal components
- Projection of data onto lower dimensional planes



Clustering using PCA



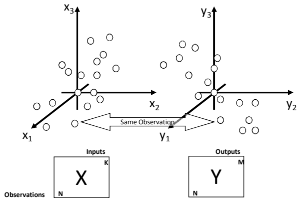
- PCA gives overview of data

7

Partial Least Squares Regression PLS

It is multivariate regression techique.

- Models structure in X and relationship to Y
- Handles
  - Correlation in both X and Y
  - Short and wide data tables
  - nVar >> nObs
  - Missing data
- Applications in
  - Spectroscopy
  - QSAR
  - Genomics Proteomics
  - ....



8

Goal of PLS regression

- The goal of PLS regression is to predict **Y from X and to describe** their common structure.
- Unlike PCA, the PLS technique works by successively extracting factors from both predictive and target variables such that **covariance** between the extracted factors is maximized.

Partial Least Squares (PLS)

Description of the technique

Assume X is a n×p matrix and Y is a nxq matrix. PLS method can work with multivariate response variables (i.e when Y is a n×q vector with q>1). However in the simplest case we can have just a single response (target).

PLS technique tries to find a linear decomposition of X and Y such that

$$X = TP^T + E$$
$$Y = TQ^T + F$$

T n×r = X-scores	U n×r = Y-scores
P p×r= X-loadings	Q 1×r = Y-loadings
E n×p = X-residual	F n×1 = Y-residual

A PLS model will try to find the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space.

Comparison of PCA and PLS

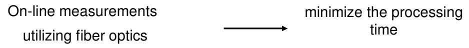
Two major common effects of using PCA or PLS

- Convert a group of correlated predictive variables to a group of independent variables
- Construct a small number of "strong" predictive variable from several "weaker" predictive variables

Major difference between PCA and PLS

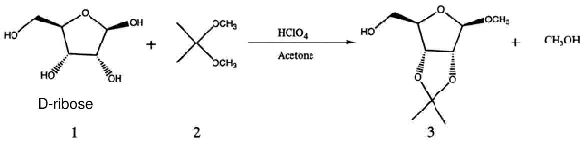
- PCA is performed without a consideration of the target variable. So PCA is an unsupervised analysis
- PLS is performed to maximized the correlation between the target variable and the predictive variables. So PLS is a supervised analysis

Prediction of methanol level using NIR spectroscopy

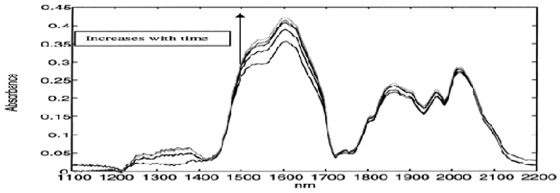

On-line measurements utilizing fiber optics
minimize the processing time

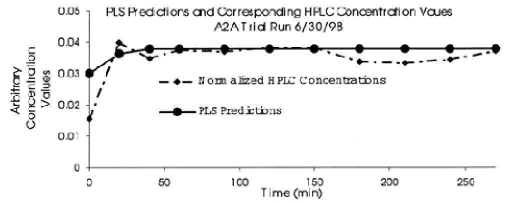
A partial least squares (PLS) calibration is built by running a number of small chemical reactions under identical conditions. Spectra are collected in real time and small aliquots are simultaneously removed from the reaction mix to perform an off-line HPLC analysis. The results of the HPLC analysis and the corresponding spectral data are input into a commercial software package, thus, creating a PLS prediction model.

S. Walker* et. Al., *Analytica Chimica Acta 395 (1999) 335-341*



Spectral analysis in the NIR is possible due to the absorbance of four O-H bonds. However, the method developed to monitor the reaction is also based on detection of methanol, a by-product which can be easily determined by NIR spectroscopy


Absorbance
Increases with time
nm


PLS Predictions and Corresponding HPLC Concentration Values
A2A Trial Run 6/30/98
Arbitrary Concentration Values
Normalized HPLC Concentrations
PLS Predictions
Time (min)