Thinking in higher dimensions and modes: Addressing multivariate multimodal multiblock data within framework of partial least-squares

Puneet Mishra[1] (*puneet.mishra@wur.nl)*

[1]Food & Biobased Research, Wageningen University & Research, The Netherlands

APACT 25 Conference

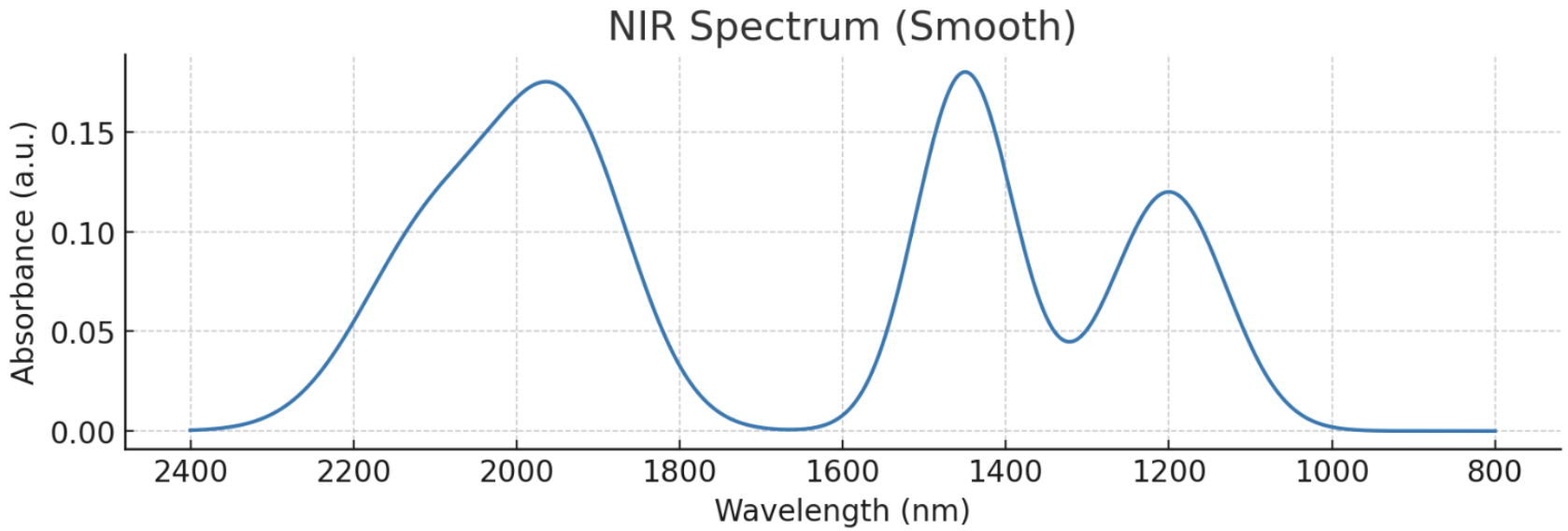
WAGENINGEN UNIVERSITY & RESEARCH

PAT sensor calibration

- Spectral measurements are performed alongside reference measurement and calibrations are made

Partial least square regression

- A statistical method that models relationships between **predictors (X)** and **response (Y)**.
- Works well when predictors are **highly correlated** or when **#predictors > #samples**.



Different PLS algorithms to approach similar solution

Research Article

Journal of CHEMOMETRICS

Received: 29 August 2008, Revised: 15 April 2009, Accepted: 8 May 2009, Published online in Wiley Interscience: 21 July 2009

(www.interscience.wiley.com) DOI: 10.1002/cem.1248

A comparison of nine PLS1 algorithms

Martin Andersson[a]*

Nine PLS1 algorithms were evaluated, primarily in terms of their numerical stability, and secondarily their speed. There were six existing algorithms: (a) NIPALS by Wold; (b) the non-orthogonalized scores algorithm by Martens; (c) Bidiag2 by Golub and Kahan; (d) SIMPLS by de Jong; (e) improved kernel PLS by Dayal; and (f) PLSF by Manne. Three new algorithms were created: (g) direct-scores PLS1 based on a new recurrent formula for the calculation of basis vectors yielding scores directly from X and y; (h) Krylov PLS1 with its regression vector defined explicitly, using only the original X and y; (i) PLSPLS1 with its regression vector recursively defined from X and the regression vectors of its previous recursions. Data from IR and NIR spectrometers applied to food, agricultural, and pharmaceutical products were used to demonstrate the numerical stability. It was found that three methods (c, f, h) create regression vectors that do not well resemble the corresponding precise PLS1 regression vectors. Because of this, their loading and score vectors were also concluded to be deviating, and their models of X and the corresponding residuals could be shown to be numerically suboptimal in a least squares sense. Methods (a, b, e, g) were the most stable. Two of them (e, g) were not only numerically stable but also much faster than methods (a, b). The fast method (d) and the moderately fast method (i) showed a tendency to become unstable at high numbers of PLS factors. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: PLS; algorithms; comparison; regression vector; numerical; stability; speed

NIPALS* algorithm

1. Information search
2. Information extraction
3. Extracted information removal from data
4. Repeat

$$for\ i = 1{:}\,a$$

$$\boldsymbol{v} = \boldsymbol{X}' * \boldsymbol{y}\ \ (Step\ 1:\ Covariance\ estimation)$$

$$\boldsymbol{t} = \boldsymbol{X} * \boldsymbol{v}\quad (Step\ 2:\ Score\ estimation)$$

$$\boldsymbol{T}(i) = \frac{\boldsymbol{t}}{norm(\boldsymbol{t})}\ (Step\ 3: Score\ normalisaiton)$$

$$\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{T}(i) * \boldsymbol{T}(i)' * \boldsymbol{X}\ (Step\ 4: Predictor\ deflation)$$

$$\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{T}(i) * \boldsymbol{T}(i)' * \boldsymbol{y}\ (Step\ 5: Response\ deflation)$$

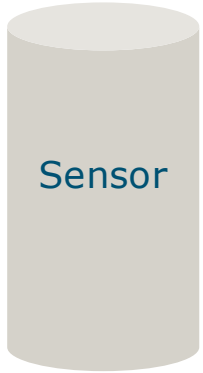
$$end$$

$\boldsymbol{X}$

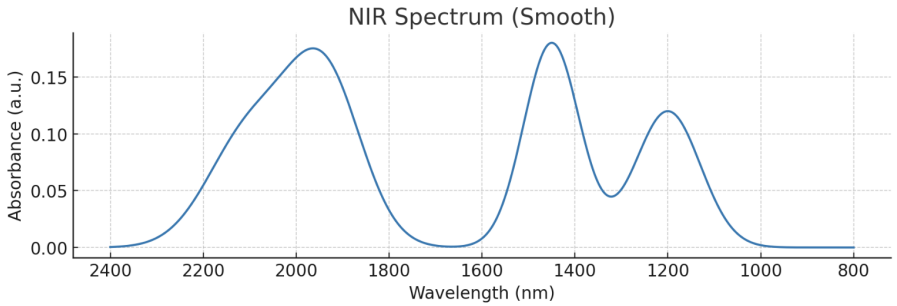
$\boldsymbol{y}$

*NIPALS : Nonlinear iterative partial least squares

Conventional use of PLS



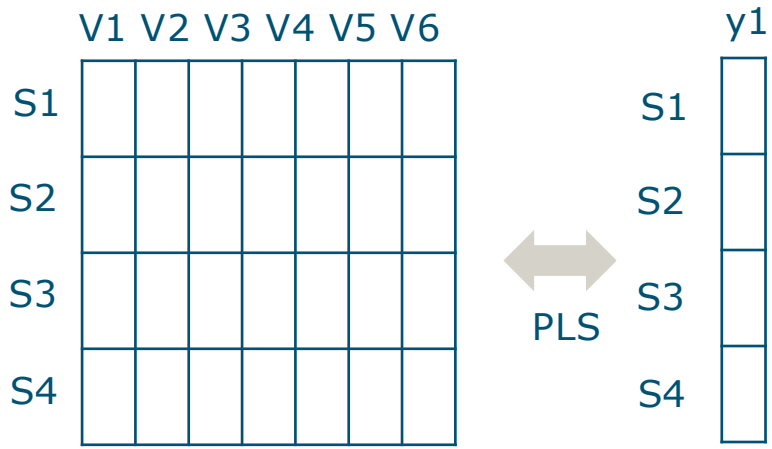
NIR, MIR, Raman etc.



Multivariate

Measurement

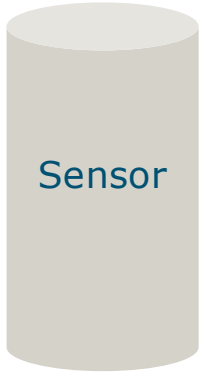
E.g. : spectra,
multiple univariates



E.g.: API concentration

Data types encountered in PAT

Types of data sets (The predictor X)

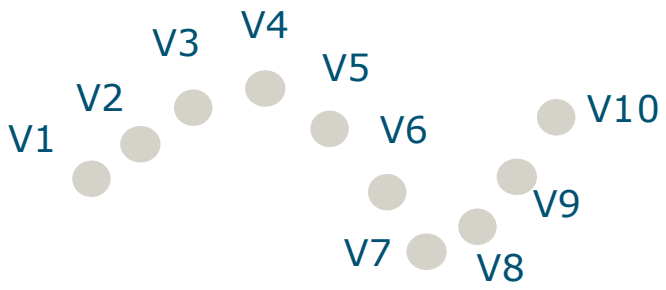


$n \times 1$

Univariate

E.g. : temperature, pH

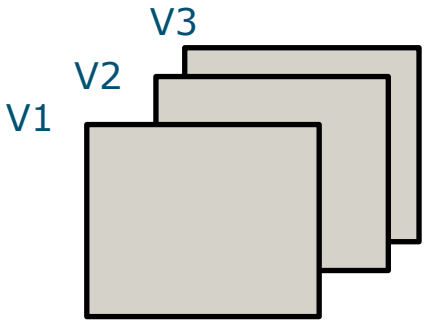
$n \times p$



Multivariate

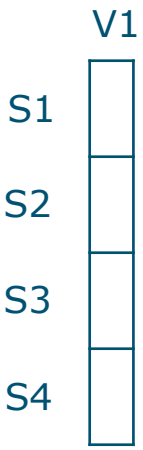
E.g. : spectra, multiple univariates

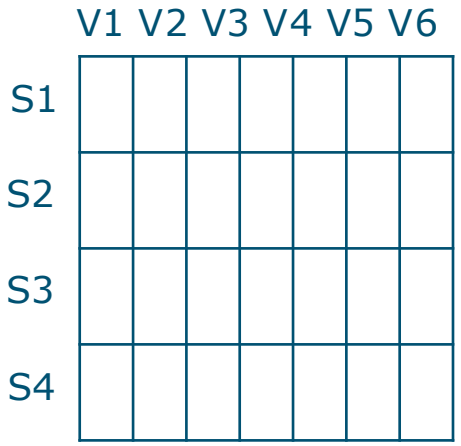
$n \times p \times k$

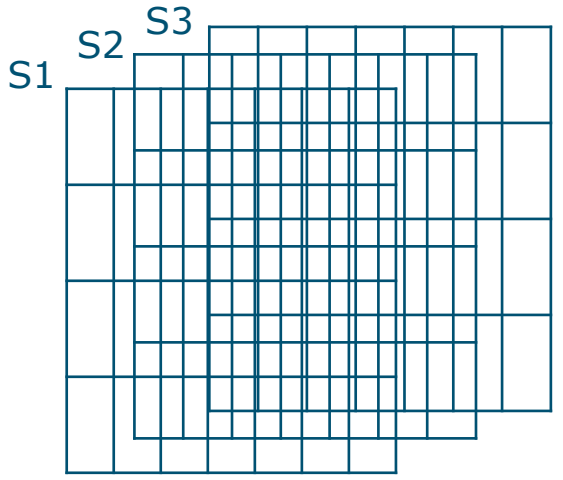


Multiway

E.g. : images, Excitation emmision floroscence, LC-GC, time series of multvariate



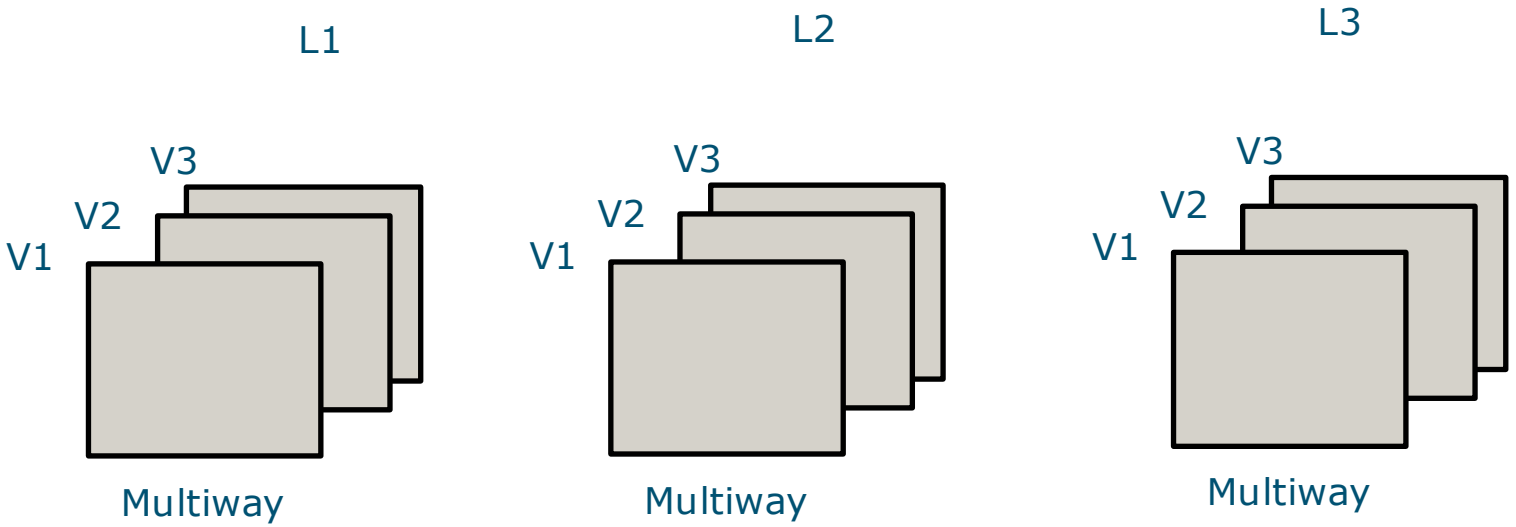




WAGENINGEN UNIVERSITY & RESEARCH

Types of data sets (The predictor X cont.)

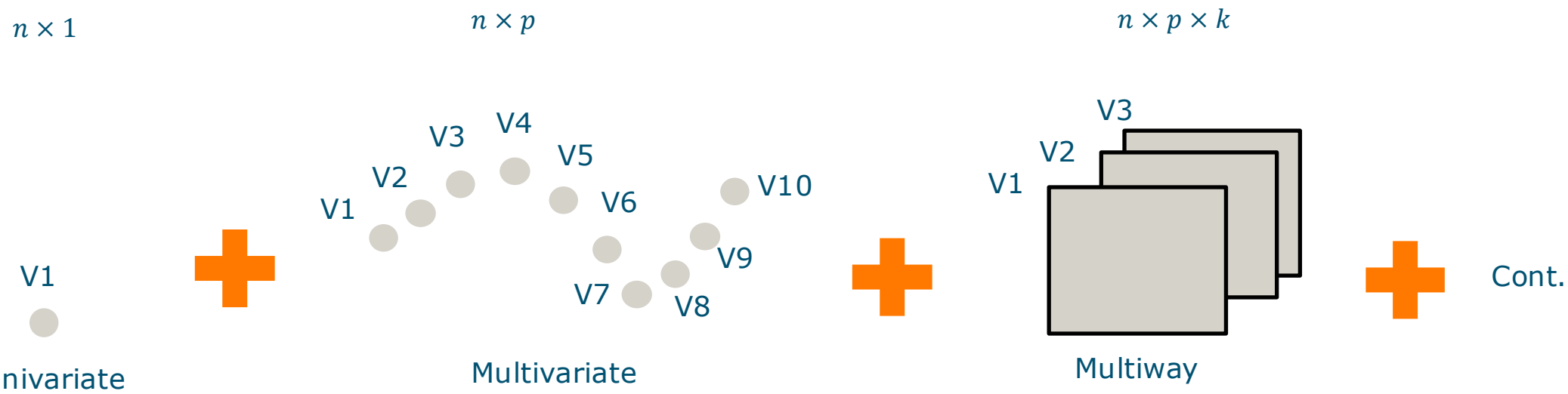
$$n \times p \times k \times l$$



E.g. : video sequences,
Excitation emmision floroscence in times,
LC-GC in time, multiple batches of time series of multivariate signal

Types of data sets (The predictor X cont.)

The Multimodal/Multisensor/Multiblock $X = \{X1, X2, X3..\}$



Univariate

E.g. : temperature, pH

Multivariate

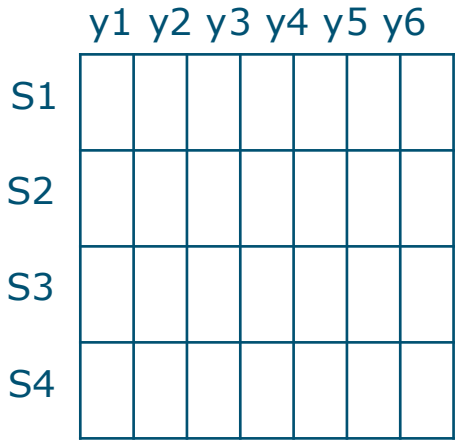
E.g. : spectra,
multiple univariates

Multiway

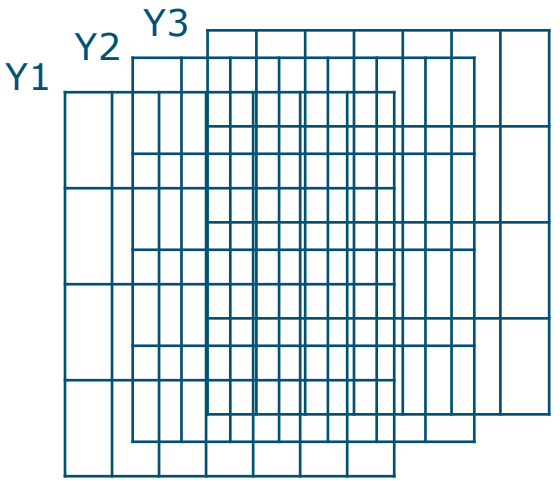
E.g. : images,
Excitation emmision floroscence,
LC-GC, time series of
multvariate

Types of data sets (The response Y cont.)

E.g.: fat concentration



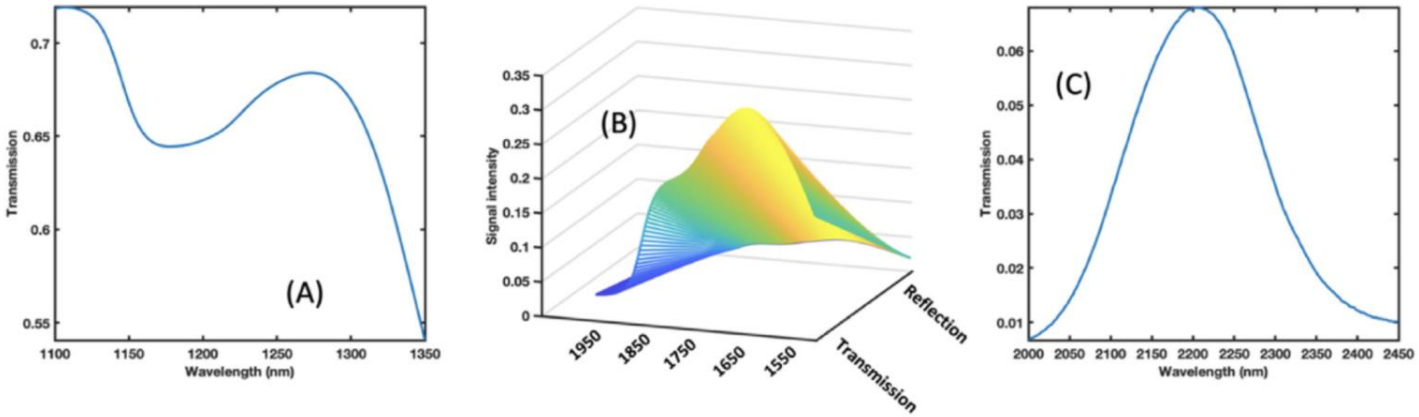
E.g.: fat and protein concentration



E.g.: fat and protein Concentration in time (maybe)

Terminology

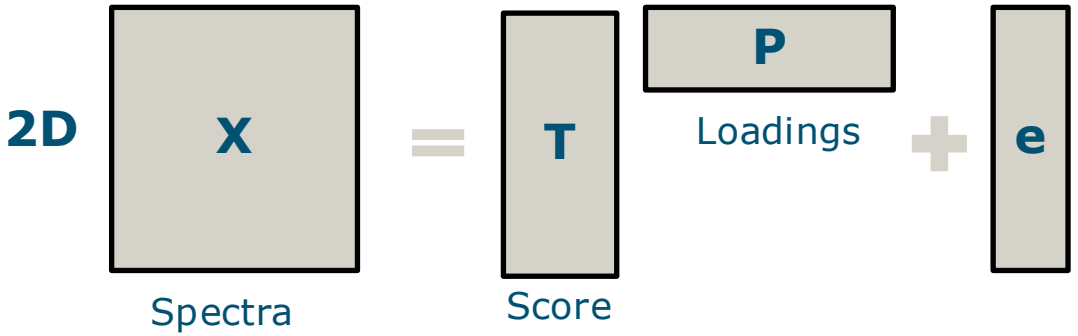
- **Multivariate** → Data with **many variables** (e.g., spectra, chromatograms, sensor arrays) measured simultaneously for each sample.
- **Multiway** → Data with **more than two modes (dimensions)** (e.g., fluorescence excitation–emission matrices, which have sample × emission × excitation).
- **Multiblock** → Data structured in **separate but related blocks** (e.g., combining spectroscopy + chromatography + sensory data for the same samples).

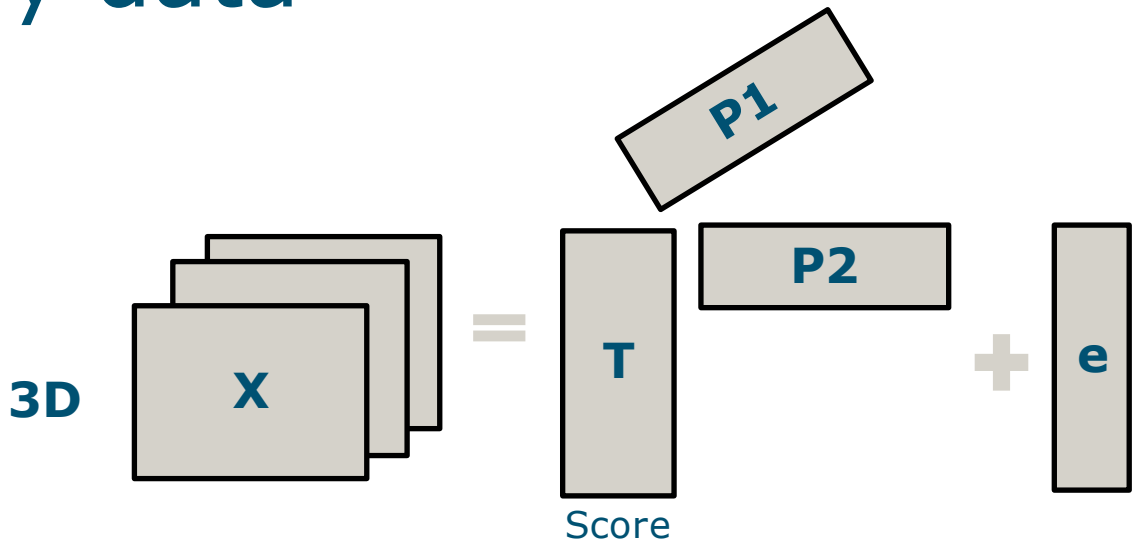


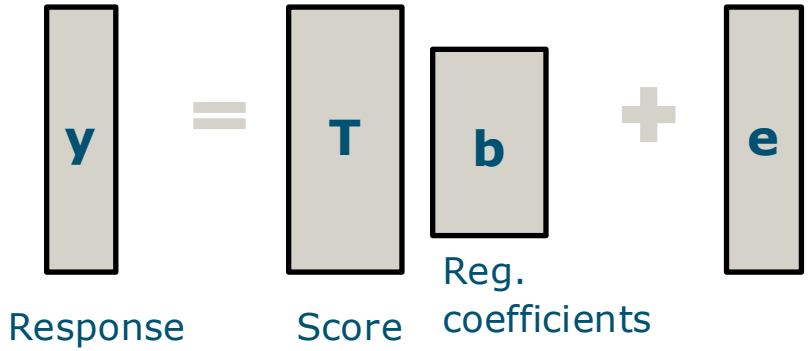
PLS algorithms depending on data types

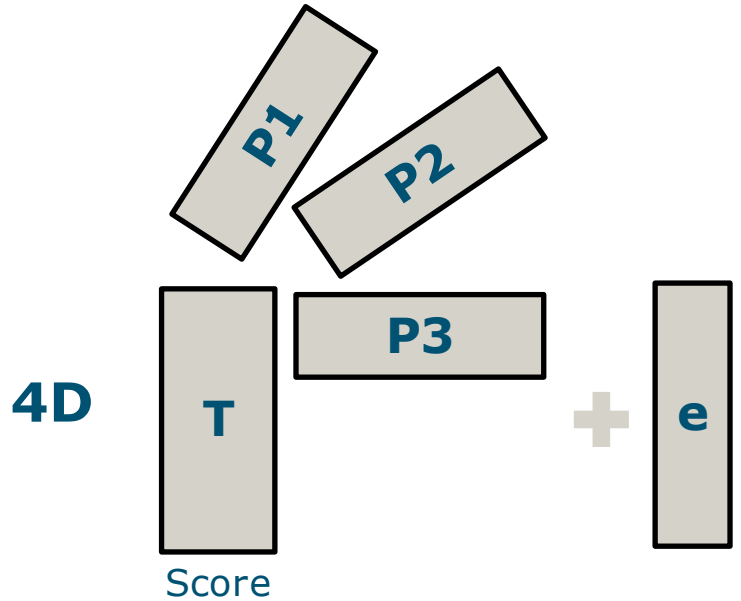
- Single two way block : **PLS**
- Single two block and multiple responses (regression and classification) : **PLS2**
- Multiple blocks : **SO-PLS, PO-PLS, MB-PLS, ROSA** etc.
- Multiway : **N-PLS, N-CPLS**
- Multiway Multiblock : **SO-N-PLS, N-ROSA, SO-N-CPLS**

How PLS handles multiway data







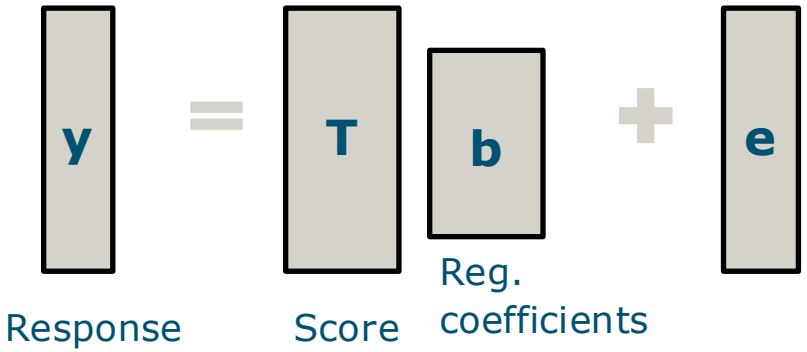


PLS aims to decompose matrices or higher order array into a set of scores and loadings.

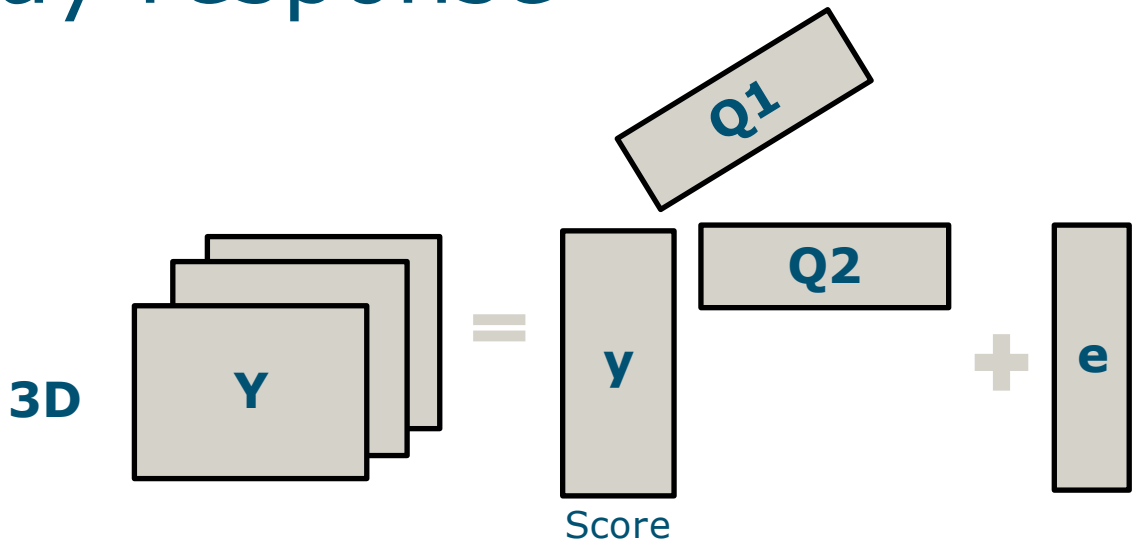
WAGENINGEN
UNIVERSITY & RESEARCH

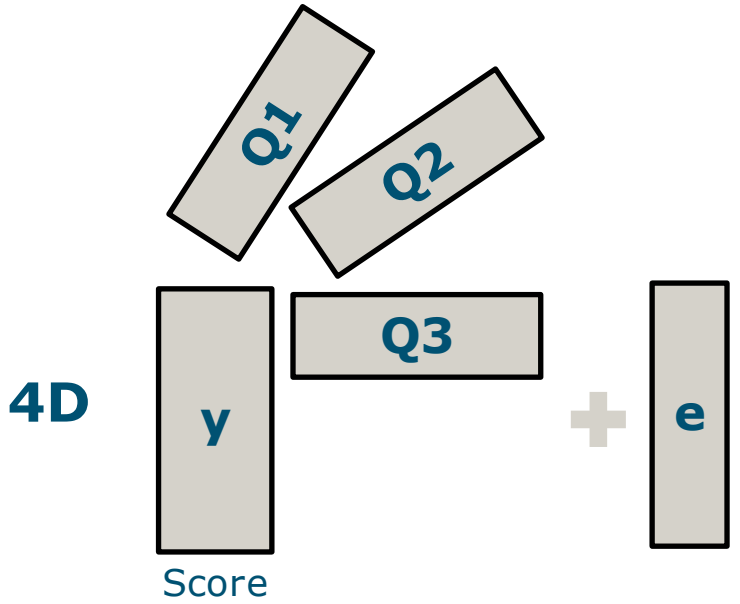
Terminology : Multilinear modelling!

How PLS handles multiway response



A response matrix or higher order array can also be decomposed into scores and loadings.



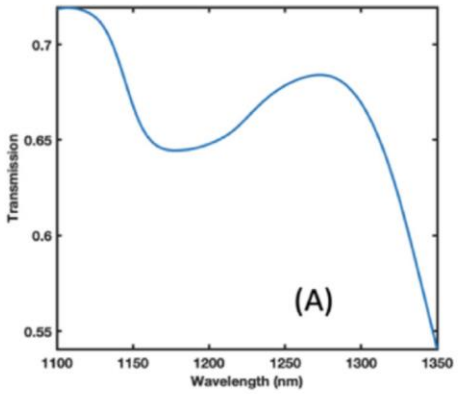


Milk multiway multiblock dataset

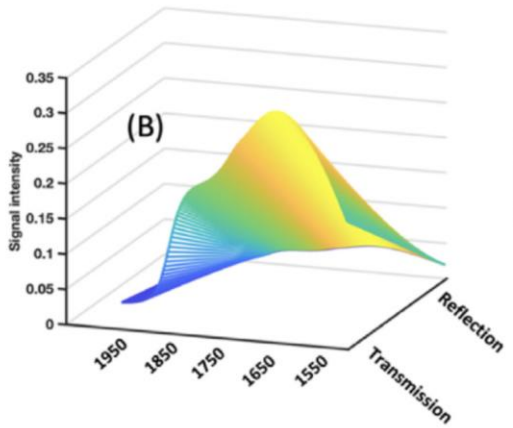
Table 1
A summary of the milk data set.

	NIRONE 1.4	NIRONE 2.0	NIRONE 2.5	Protein (% w/w)	Fat (% w/w)
Spectral range (nm)	1100–1350	1550–1950	2000–2450	*	*
Data shape	296 × 126	296 × 201 × 2	296 × 226	296 × 1	296 × 1
Reference range (Average ± standard deviation)	*	*	*	3.90 ± 0.41	4.71 ± 1.10

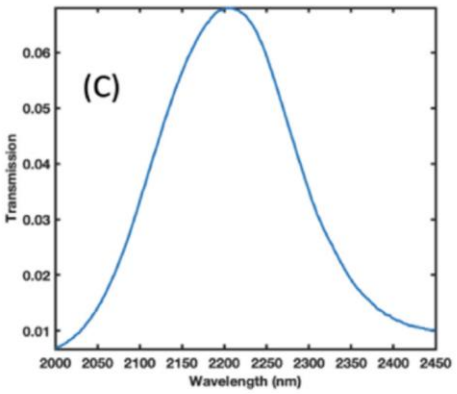
*not relevant.



Block 1



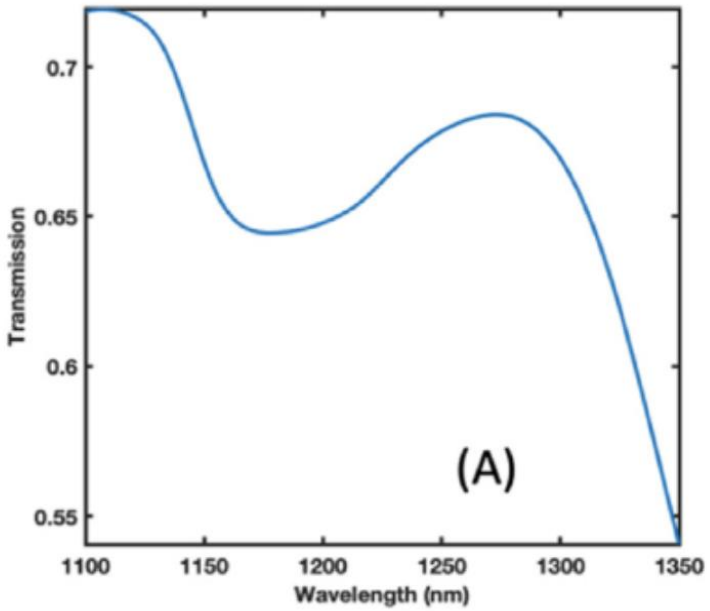
Block 2

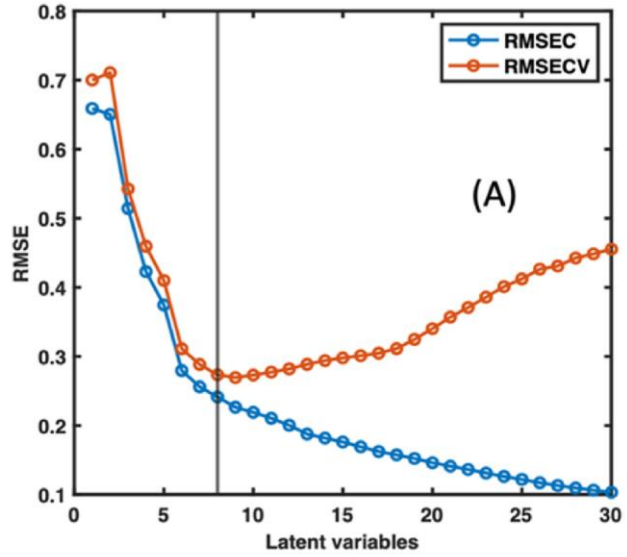


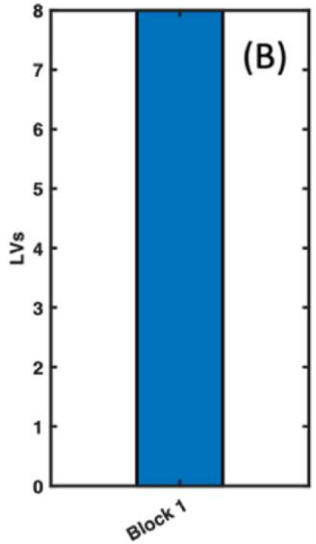
Block 3

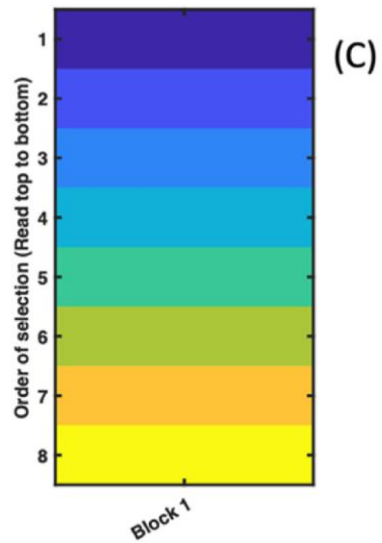
WAGENINGEN
UNIVERSITY & RESEARCH

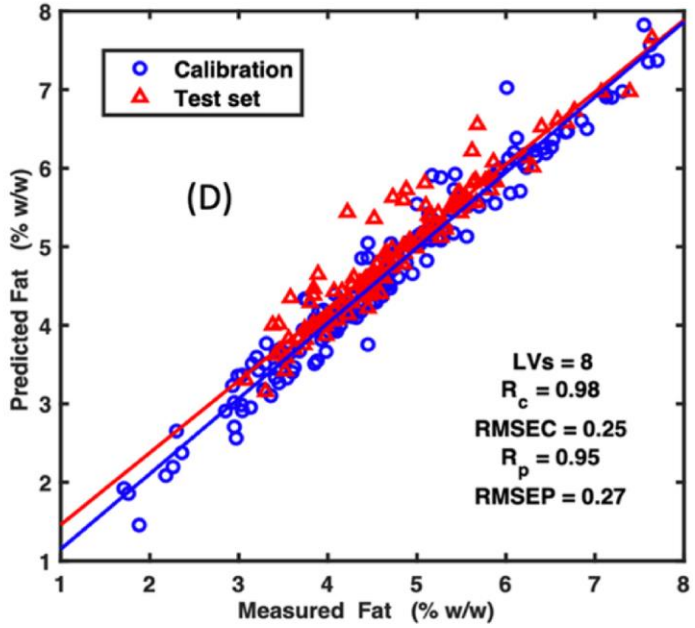
PLS analysis of single two-way data





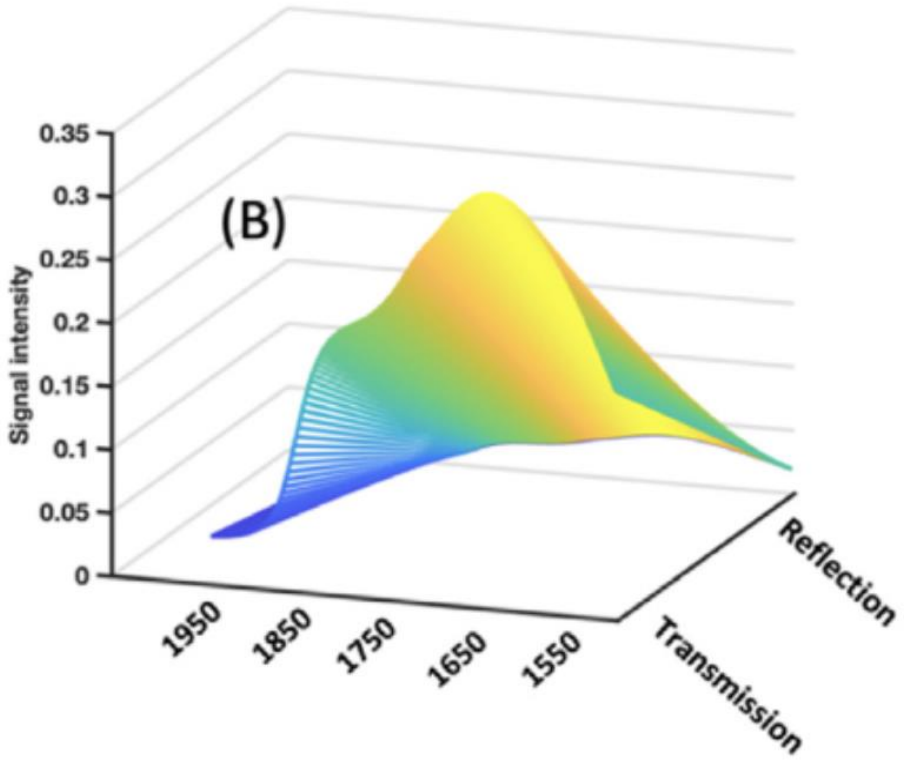


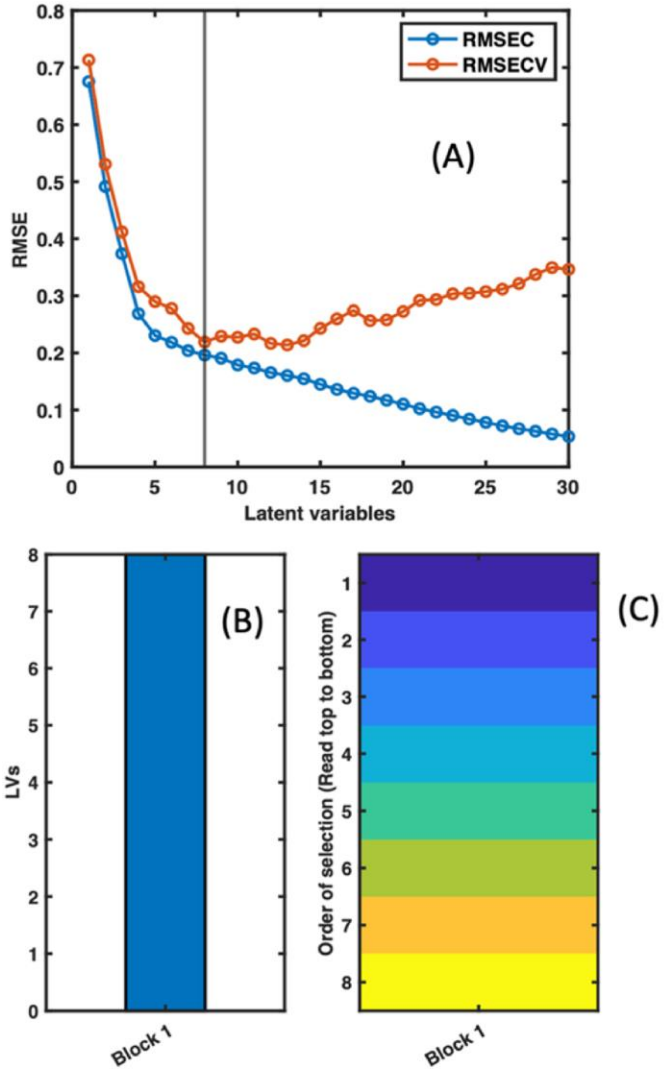


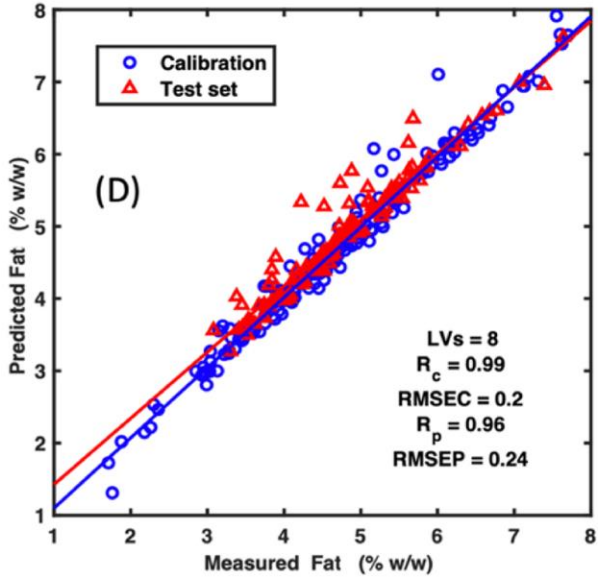


WAGENINGEN
UNIVERSITY & RESEARCH

N-PLS analysis of multiway data

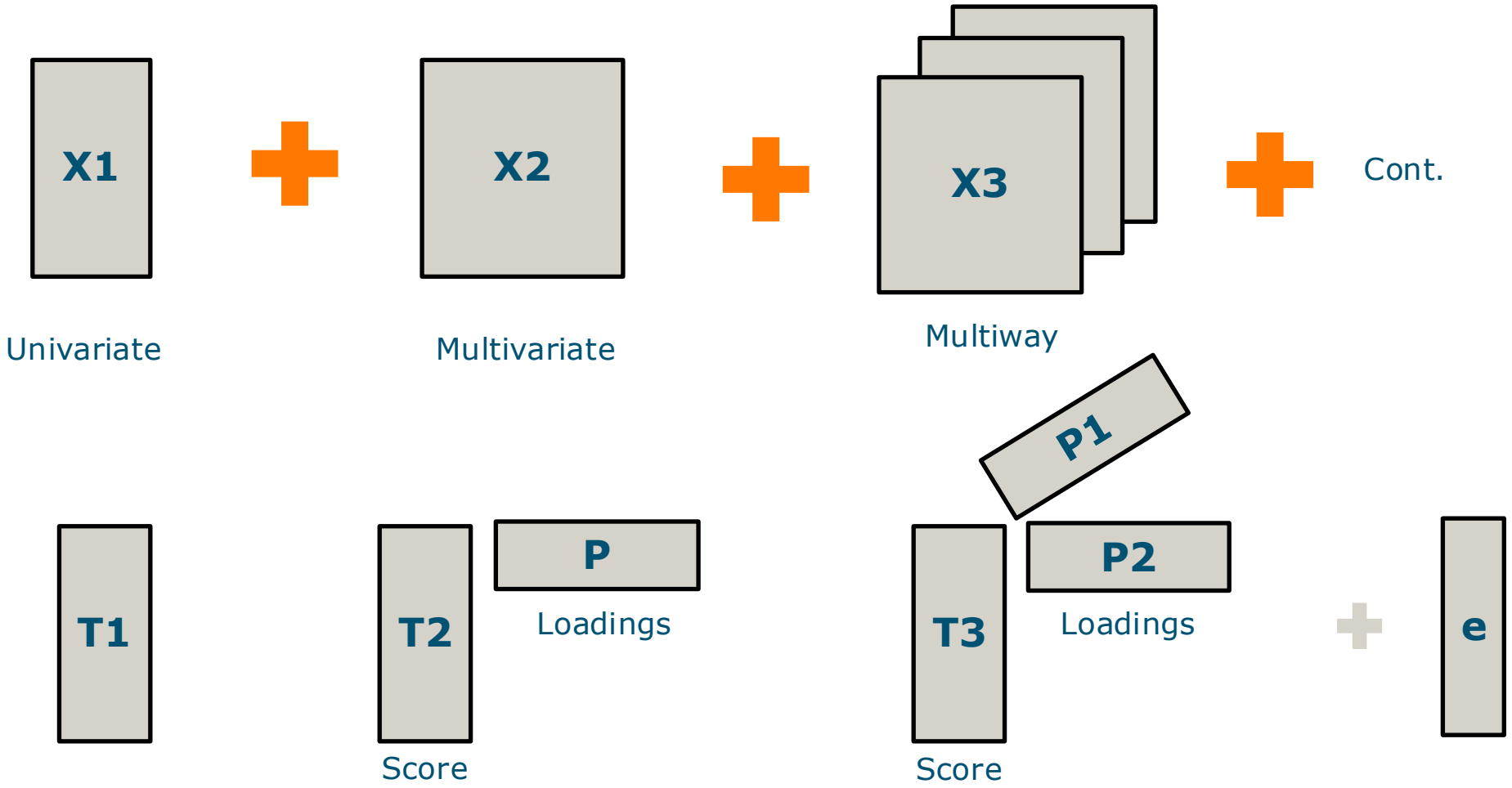




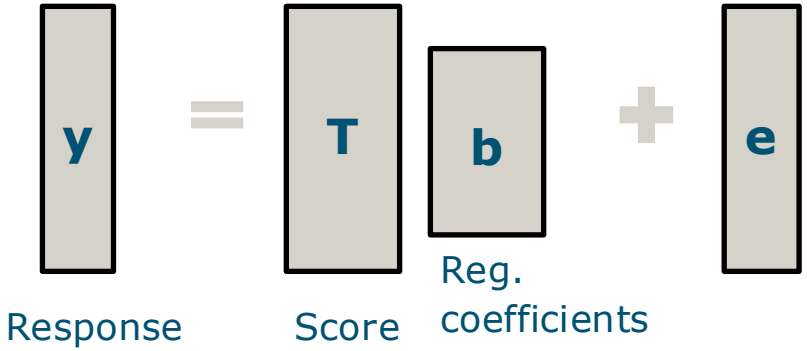


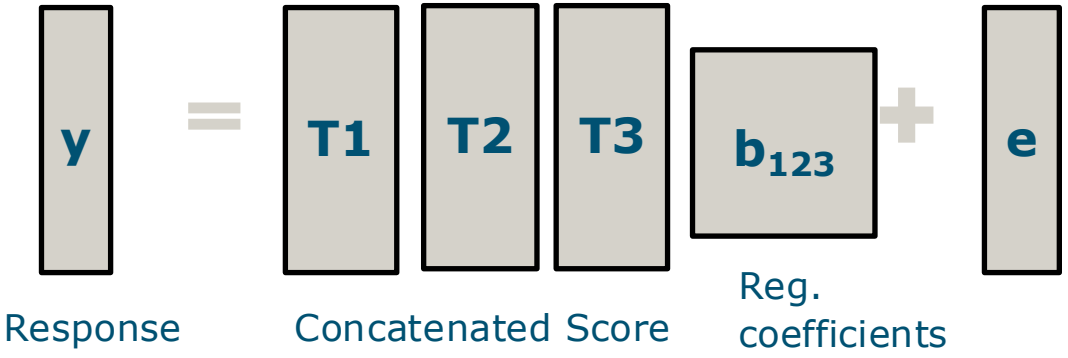
How PLS handles multiblock data

The Multimodal/Multisensor/Multiblock $X = \{X1, X2, X3..\}$



How PLS handles multiblock data





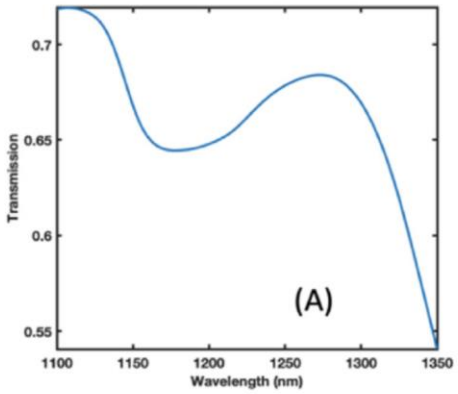
Scale independent!

Milk multiway multiblock dataset

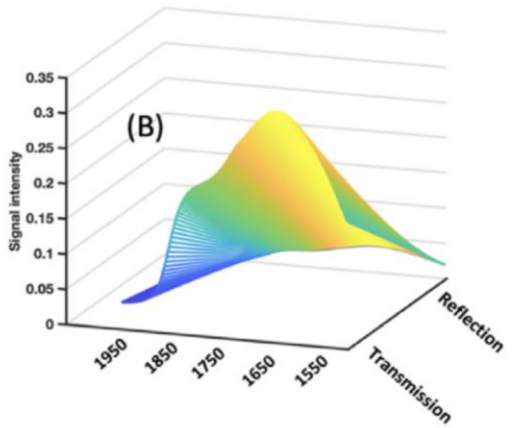
Table 1
A summary of the milk data set.

	NIRONE 1.4	NIRONE 2.0	NIRONE 2.5	Protein (% w/w)	Fat (% w/w)
Spectral range (nm)	1100–1350	1550–1950	2000–2450	*	*
Data shape	296 × 126	296 × 201 × 2	296 × 226	296 × 1	296 × 1
Reference range (Average ± standard deviation)	*	*	*	3.90 ± 0.41	4.71 ± 1.10

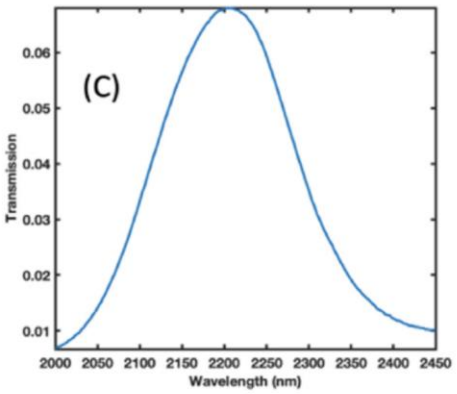
*not relevant.



Block 1

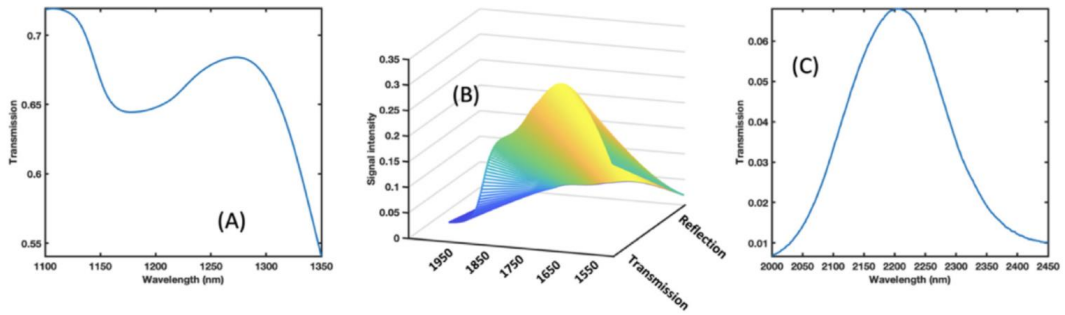


Block 2

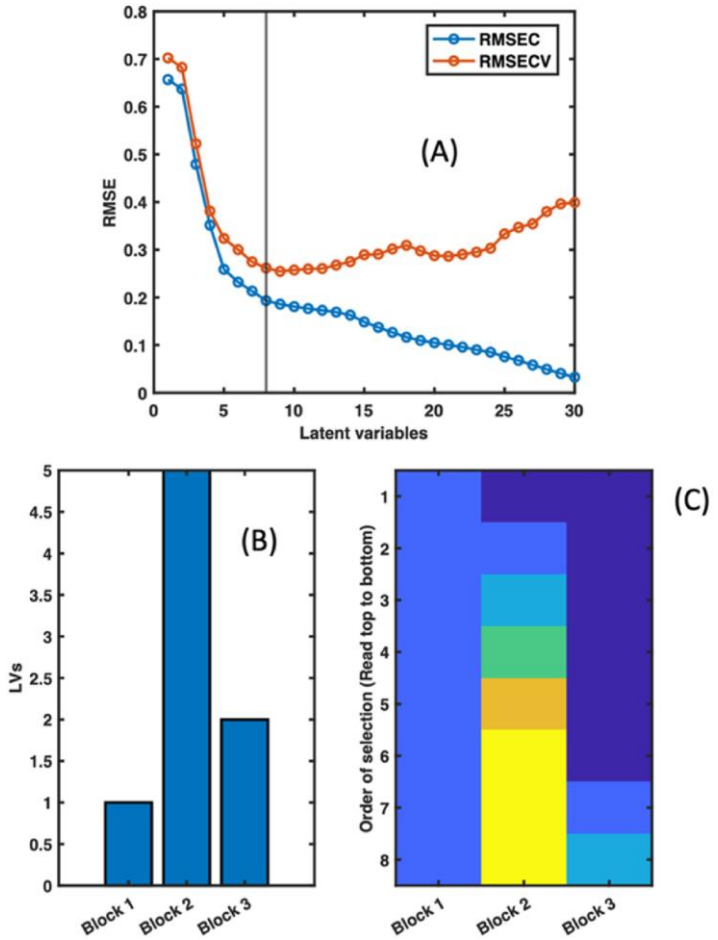


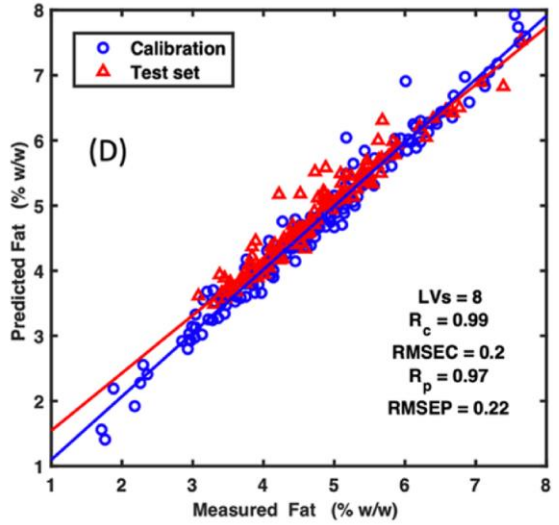
Block 3

Multiblock multiway modelling



Complementary information is learned from different data blocks.

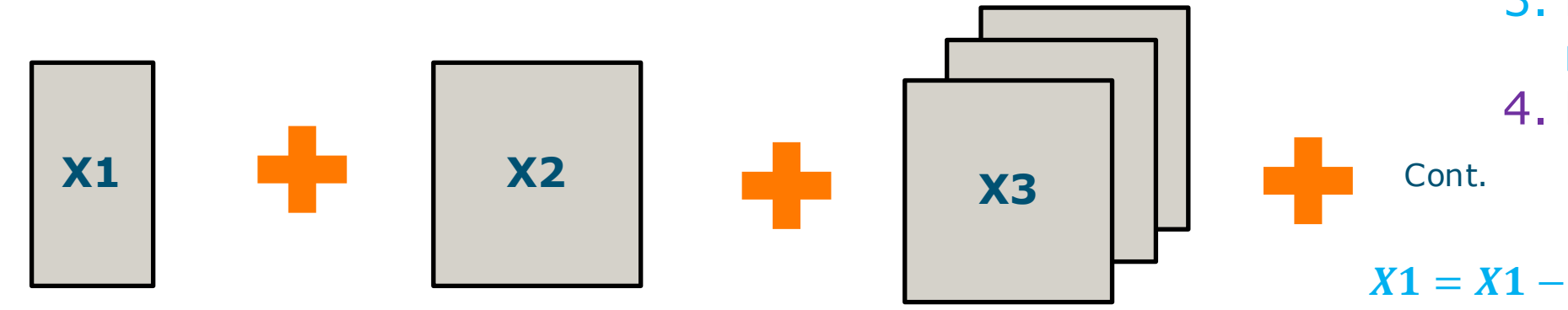




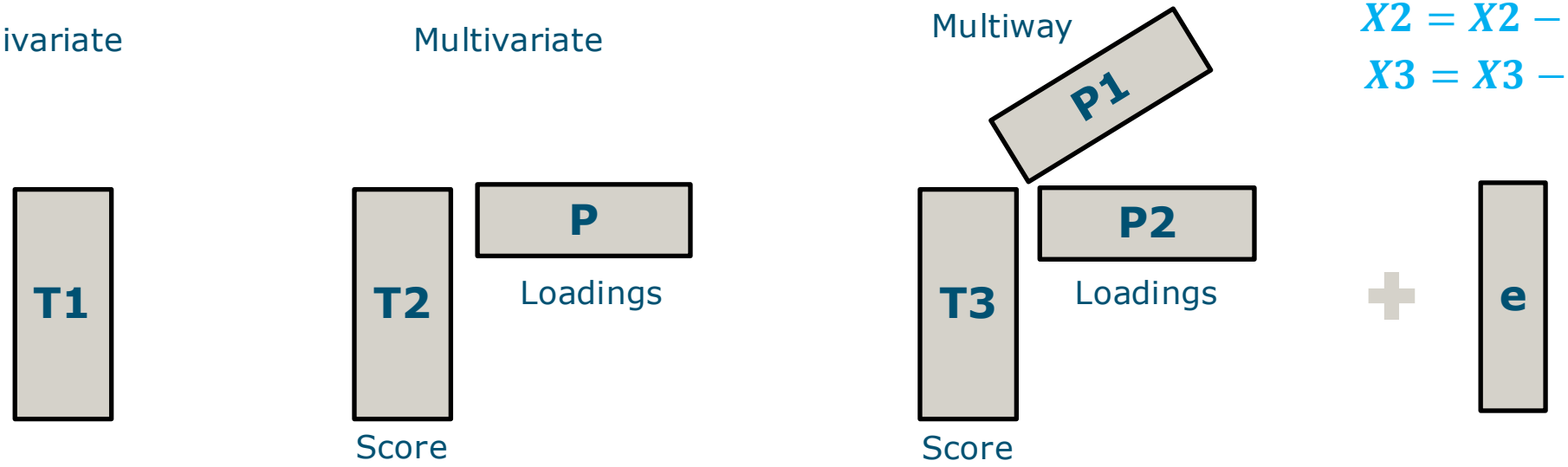
How PLS models complementary information

The Multimodal/Multisensor/Multiblock $X = \{X1, X2, X3..\}$

1. Information search
2. Information extraction
3. Extracted information removal from data
4. Repeat

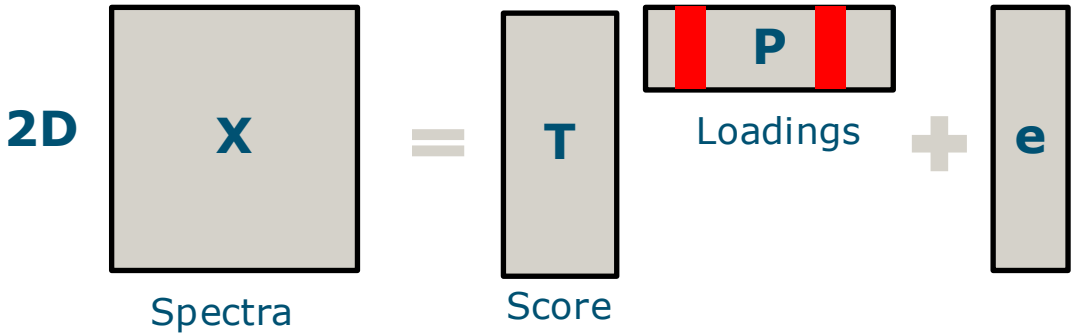


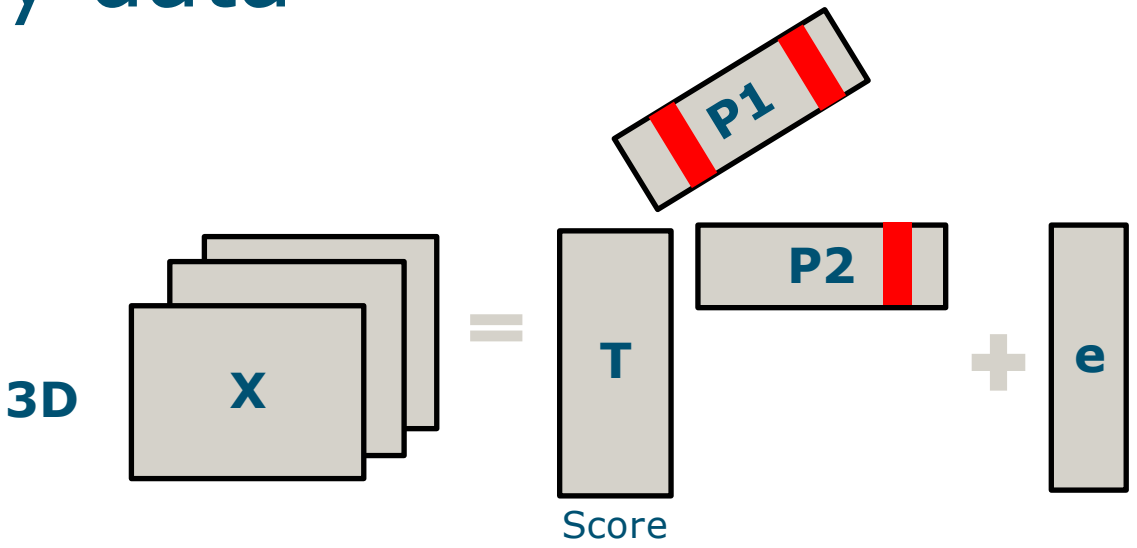
$$\boldsymbol{X1} = \boldsymbol{X1} - \boldsymbol{T}(i) * \boldsymbol{T}(i)' * \boldsymbol{X1}\ (deflation)$$
$$\boldsymbol{X2} = \boldsymbol{X2} - \boldsymbol{T}(i) * \boldsymbol{T}(i)' * \boldsymbol{X2}\ (deflation)$$
$$\boldsymbol{X3} = \boldsymbol{X3} - \boldsymbol{T}(i) * \boldsymbol{T}(i)' * \boldsymbol{X3}\ (deflation)$$

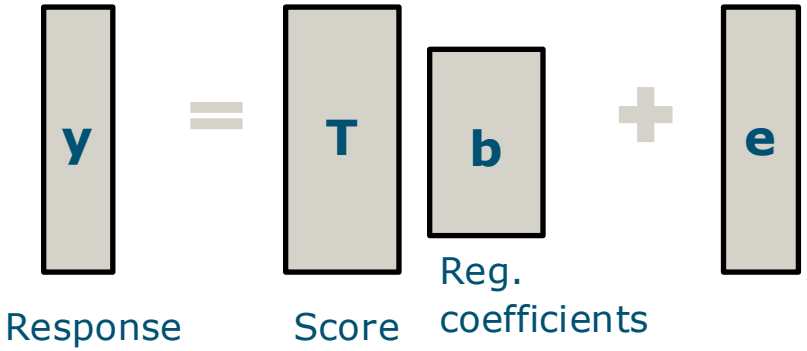


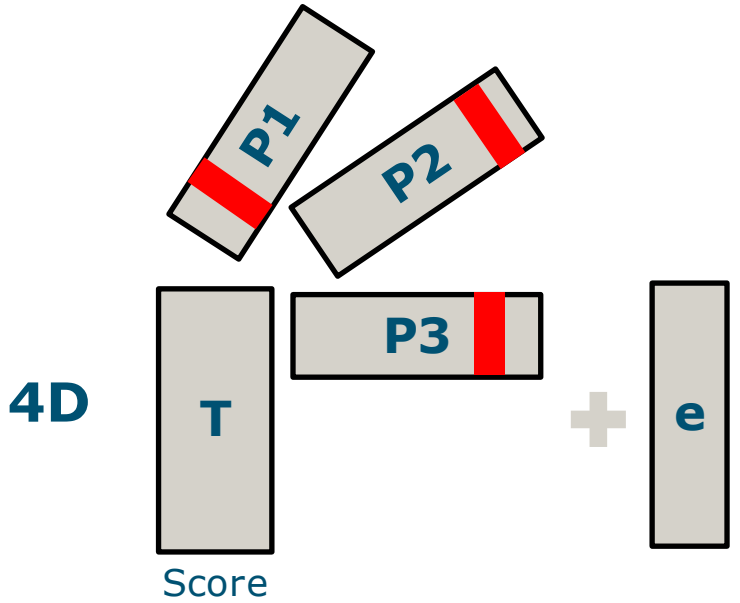
How to select features in multiway, multiblock data

Feature selection multiway data



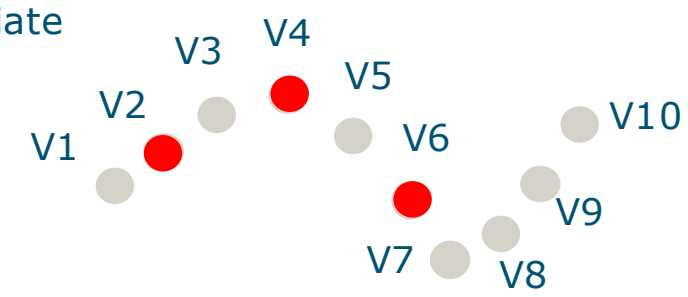




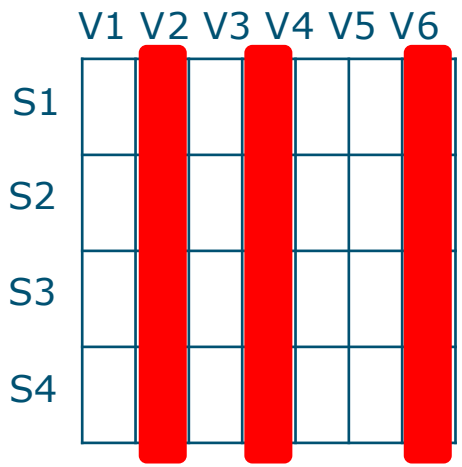


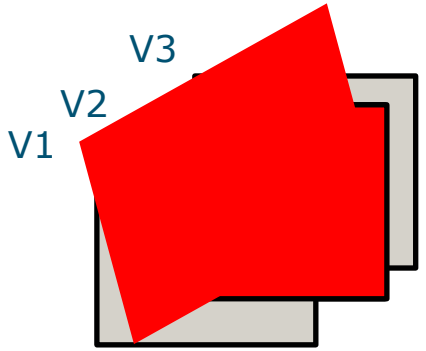
Feature selection : identifying important variables

Multivariate



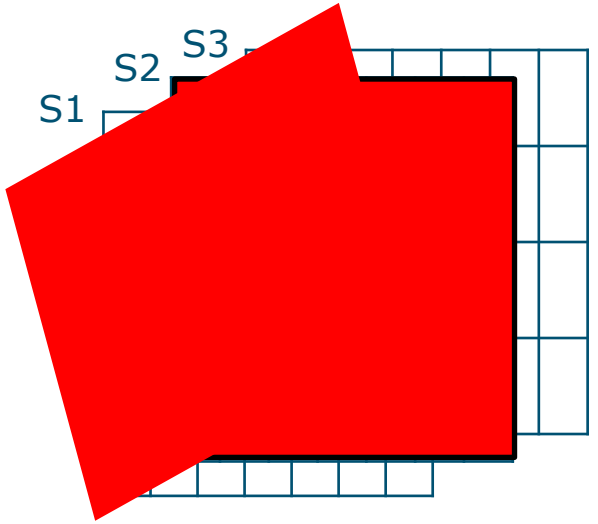
E.g. : spectra, multiple univariates



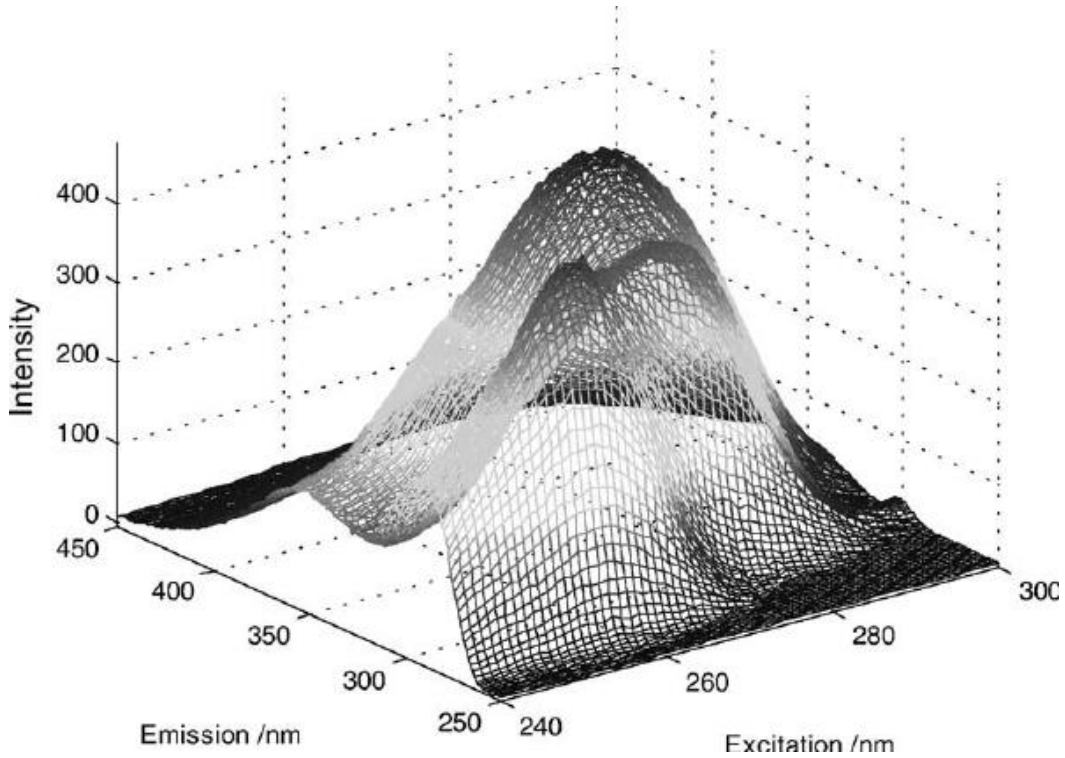


Multiway

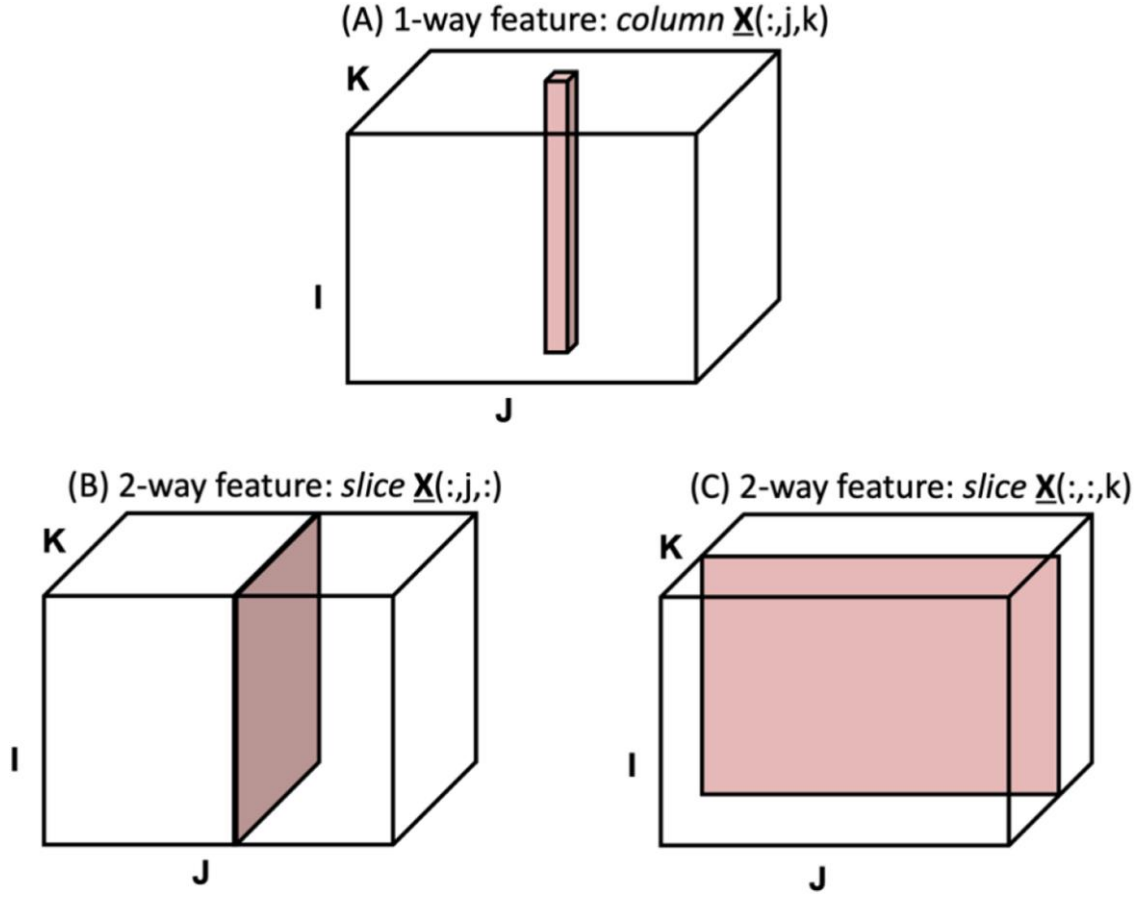
E.g. : images,
Excitation emmision floroscence, LC-GC, time series of multvariate



Intutive example for higher dimensional features

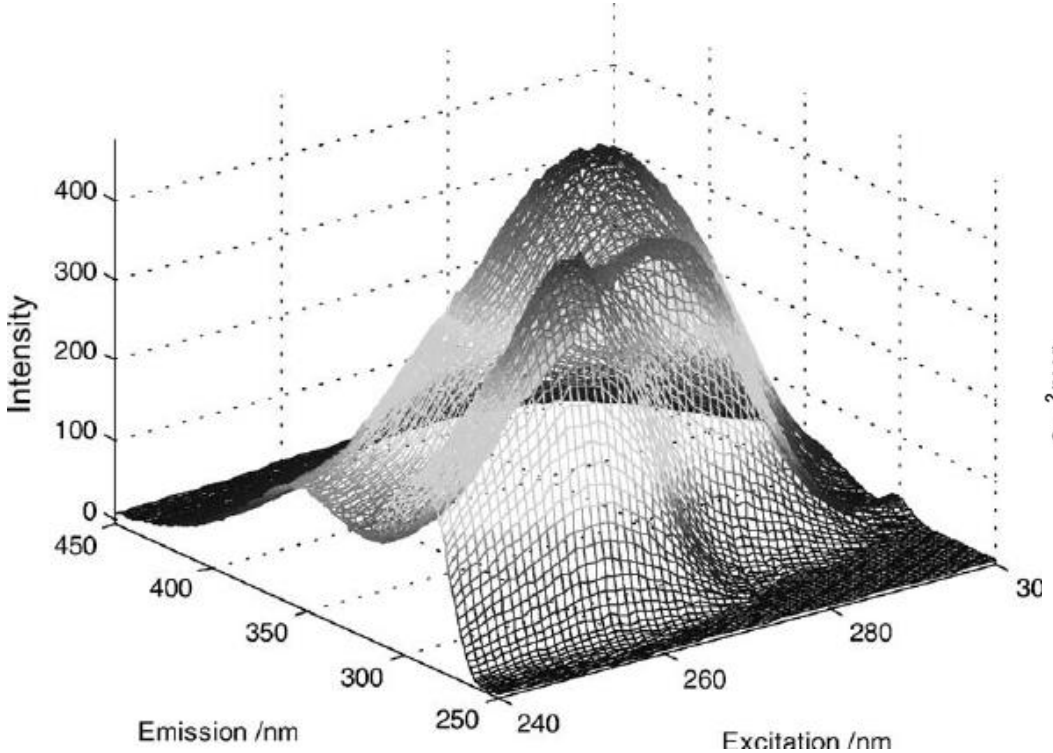


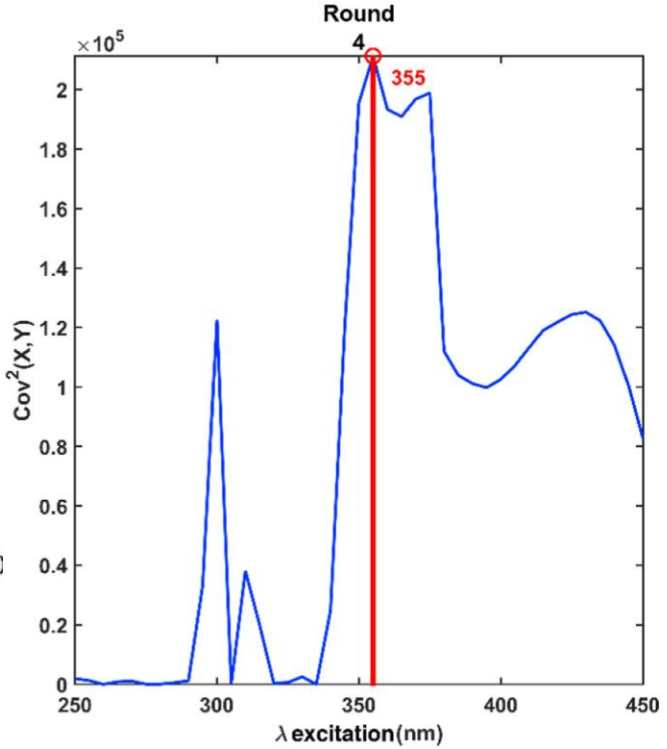
Fouroscenece excitation emission data

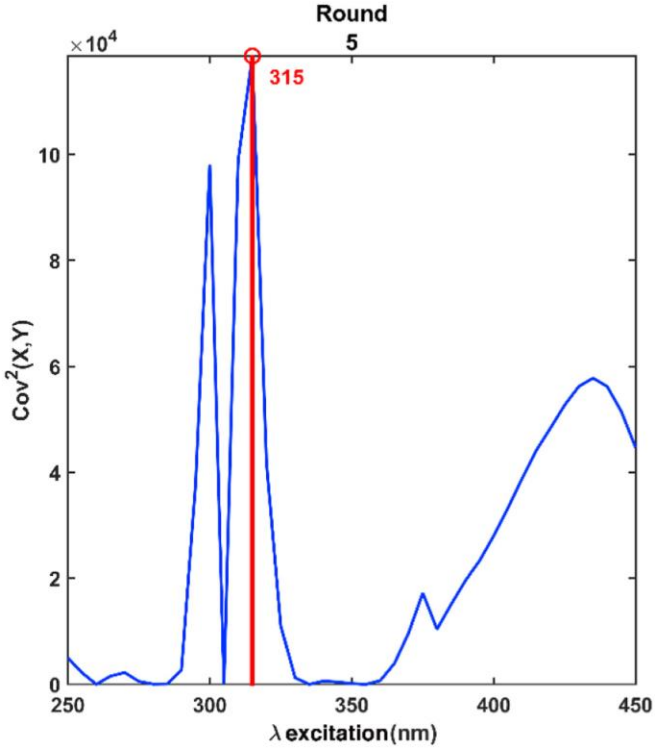


Biancolillo, Alessandra, Jean-Michel Roger, and Federico Marini. "N-CovSel, a new strategy for feature selection in N-way data." *Analytica Chimica Acta* 1231 (2022): 340433.

Intuitive example for higher dimensional features



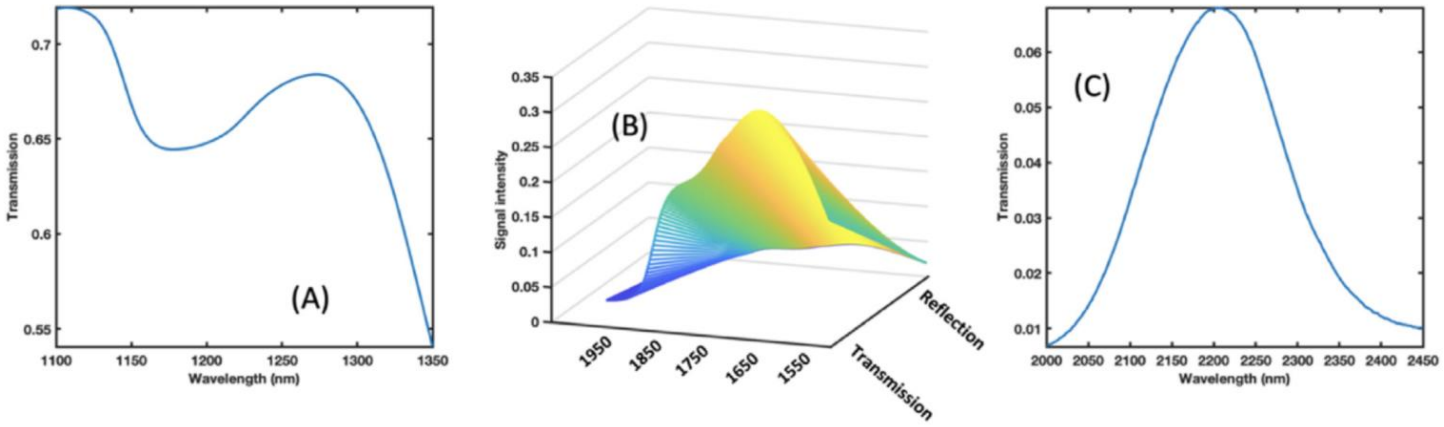




Fouroscence excitation emission data

"For any n-way data, features can be selected of type 1D to n-1 way"

Multiway multiblock feature selection modelling



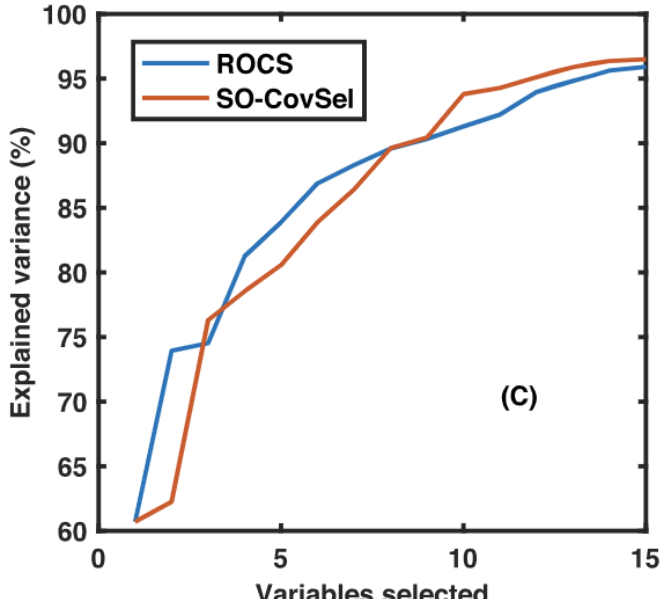
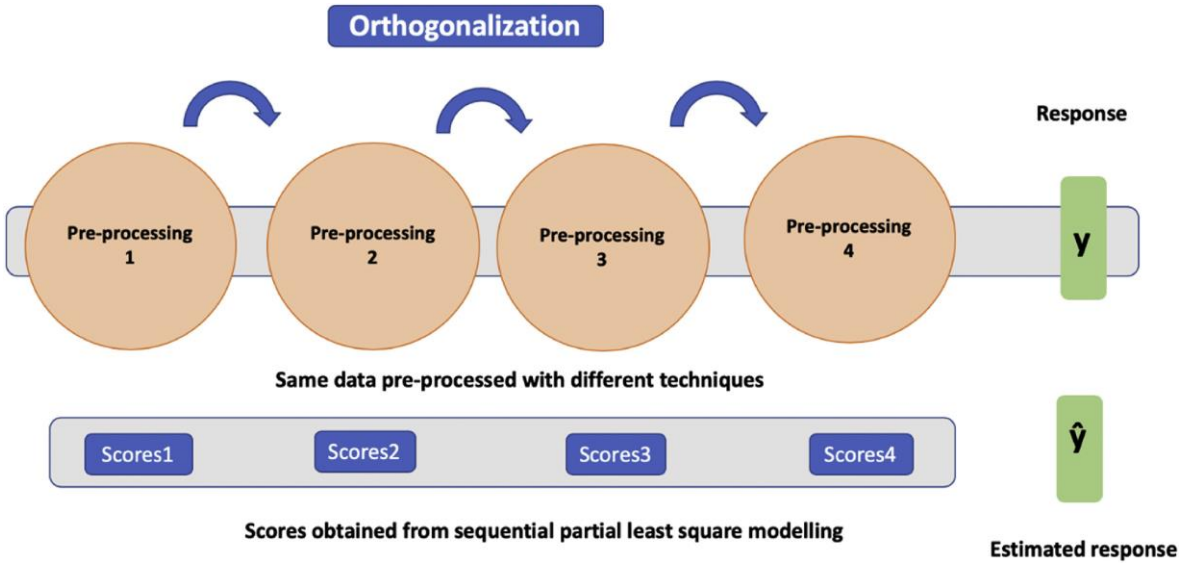
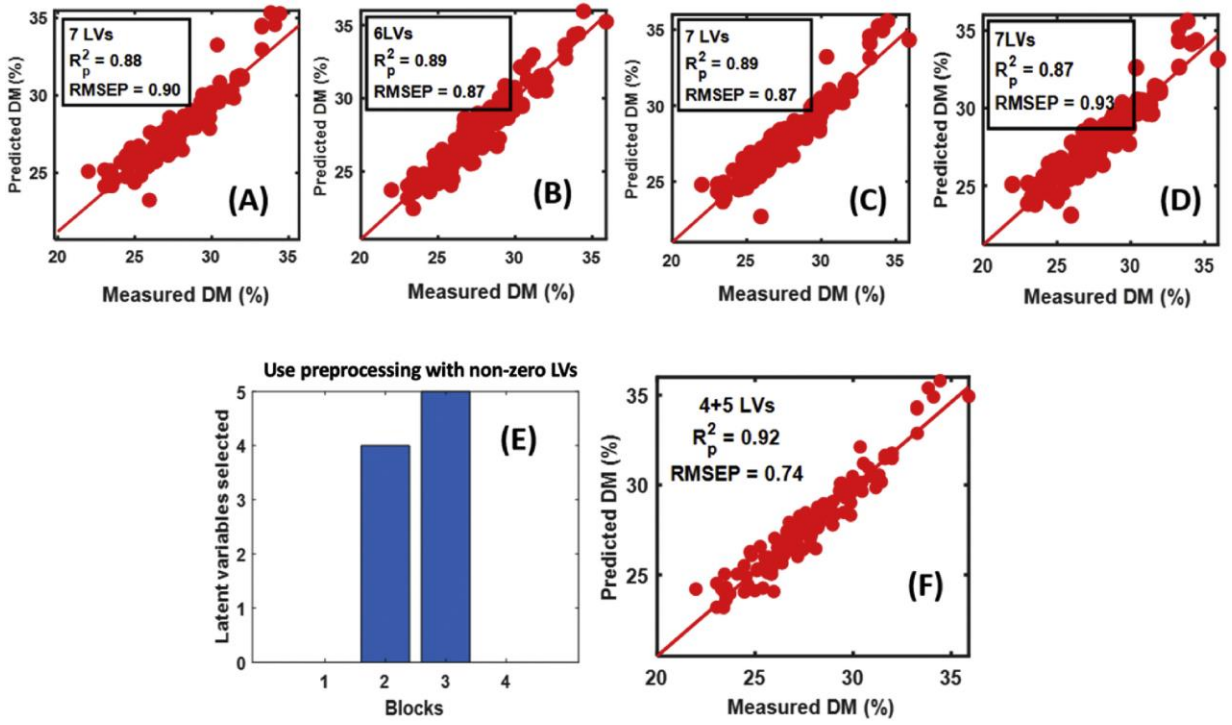


TABLE 4 A summary of features selected for multiblock multiway data

Method	First block	Second block	Third block
ROCS	1296, 1350	(1666,2), (1722,1), (1608,1), (1838,1), (1844,2)	2110, 2298, 2180, 2128, 2244, 2196, 2290

An example of pre-processing fusion





Data pre-processed with different pre-processing can be treated as multiblock data!

Mishra, Puneet, et al. "New data preprocessing trends based on ensemble of multiple preprocessing techniques." *TrAC Trends in Analytical Chemistry* 132 (2020): 116045.

Two unified algorithms for multiway multiblock modelling and feature selection

Analytica Chimica Acta
Volume 1206, 8 May 2022, 339786

Swiss knife partial least squares (SKPLS): One tool for modelling single block, multiblock, multiway, multiway multiblock including multi-responses and meta information under the ROSA framework

Puneet Mishra [a], Kristian Hovde Liland [b]

Journal of CHEMOMETRICS

RESEARCH ARTICLE | Open Access

Swiss knife covariates selection: A unified algorithm for covariates selection in single block, multiblock, multiway, multiway multiblock cases including multiple responses

Puneet Mishra, Kristian Hovde Liland, Ulf Geir Indahl

First published: 02 September 2022 | https://doi.org/10.1002/cem.3441 | Citations: 2

Pre-recorded CPACT webinars

CPACT

WAGENINGEN UNIVERSITY & RESEARCH

Conclusions

- PLS is a versatile framework which allows handling wide variety of data types in PAT.
- Its unique property to perform multilinear modelling allows handling 1D to nD data.
- It allows a scale independent data fusion framework to model complementary information.
- Feature selection in multiway data is also possible within PLS framework.
- Several uncommon extensions of PLS can benefit PAT data processing.

Free (Gratis) training

Training

Chemometric Approaches for Hyperspectral Image Processing

Date **19-21 November 2025,** each day from 9:00 to 17:00 hr

Location Phenomea, building number 125 on **Wageningen Campus**

Course leader Dr. P (Puneet) Mishra, Wageningen Food & Biobased Research

Day 2 & 3, Thu. 20 and Fri. 21 November 2025

Theme: Open-Source Python Tutorial Session - Mastering Chemometrics

- **Hyperspectral Image Loading, Pre-processing, Object Detection/Segmentation**
  - Efficiently handle and prepare your hyperspectral cubes.
  - Techniques for identifying and isolating regions of interest.
- **Spectral Extraction, Spectral Pre-processing, Outliers Removal, Chemometric Exploratory & Predictive Modelling**
  - Extracting meaningful spectral signatures.
  - Noise reduction, baseline correction, and scatter correction.
  - Unveiling patterns with PCA, and building robust prediction models (e.g., PLSR, PCR, deep learning).
- **Variable Selection**
  - Identifying the most informative wavelengths for model efficiency and interpretability.
- **Robust Modelling**
  - Strategies for building models that perform reliably across varying conditions.
- **Fusion of Information from Different Sensors**
  - Integrating data from multiple sources for comprehensive analysis.
- **Model Transfer Between Sensors**
  - Techniques for applying models developed on one sensor to another.
- **Model Robustness**
  - Assessing and enhancing the stability and reliability of your models.
- **Model Maintenance and Correction Strategies**
  - Ensuring long-term performance and adaptability of deployed models.
- **Using pre-trained AI models for hyperspectral image processing**
  - How to use already trained open-source AI models to improve hyperspectral modelling