Comparison of chemometrics methods for the spectroscopic monitoring of chemical reactions

04/10/2018

Michel Thiel, Bernadette Govaerts

Manufacturing and Applied Statistics (MAS)

janssen | PHARMACEUTICAL COMPANIES OF Johnson&Johnson

Context

Collaboration with chemists

- Transfer of chemical reactions from lab to plant scale

Chemical reactions monitored by spectroscopy and chromatography

- Complex data

Helping chemists with statistics

- Better understanding of chemical reactions
- Better control of the quality

Comparison of chemometrics methods

Overview

Data and objectives

Statistical methods

Visualization and interpretation of chemical reactions

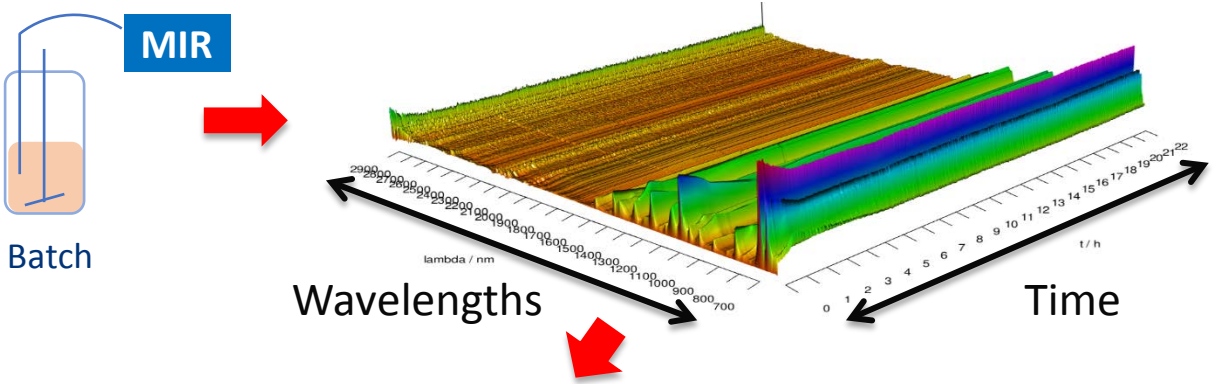
- Dimension-reduction methods

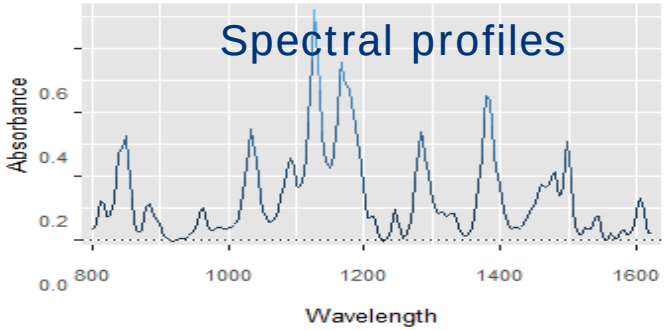
Prediction of chemical concentrations

- Calibration methods

Data and objectives

Mid-infrared (MIR) spectroscopy





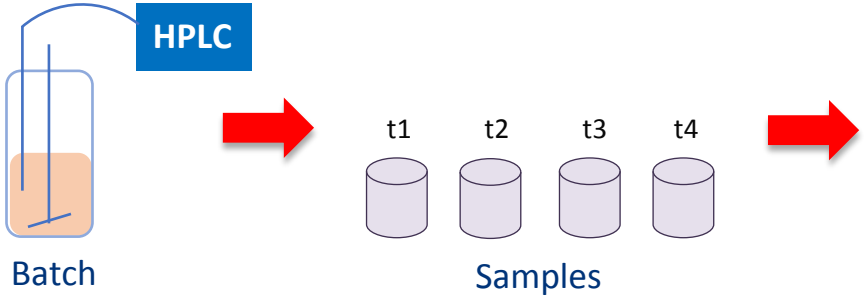
Chemical reaction monitored over time

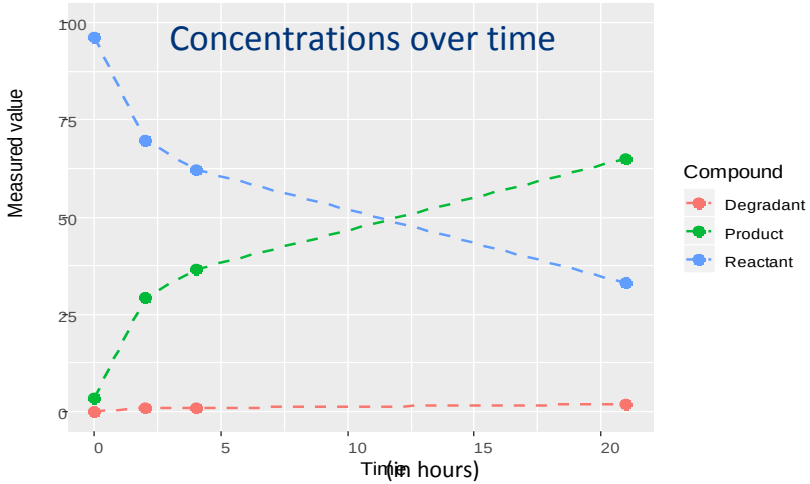
>1000 spectra over >20 hours

Continuous monitoring

Complex data

High-Pressure Liquid Chromatography (HPLC)





Samples taken over the reaction

Separation of molecules by chromatography

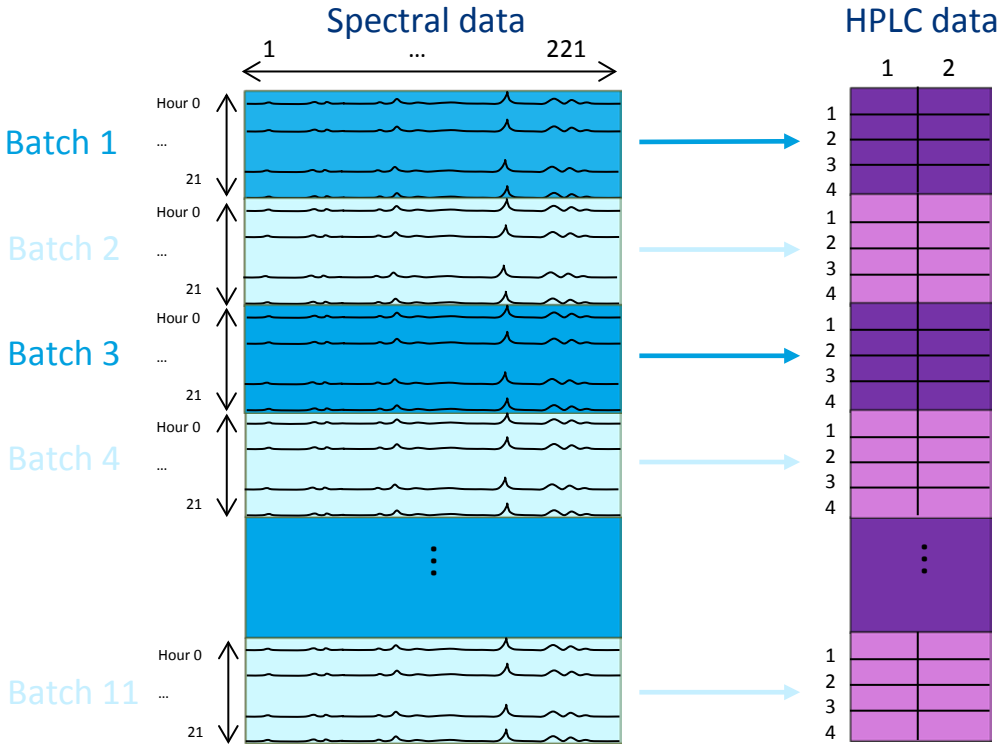
HPLC data << Spectroscopic data

Manual sampling

Quantification by a UV detector

Concentrations of molecules

Datasets



11 batches

- MIR and HPLC monitoring
- Synthesis of the same compound

Spectral matrix $X_{T\times m}$

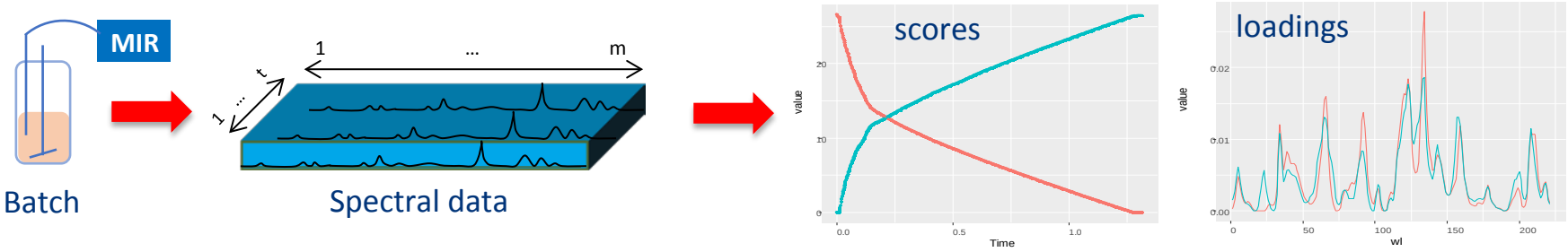
- 11 reactions
- Over 21 hours
- Dimension 6510×211

HPLC matrix $Y_{S\times l}$

- 11 reactions
- Final product and reactant
- 0, 2, 4 and 21 hours
- Dimension 44×2

Objective 1: visualization and interpretation

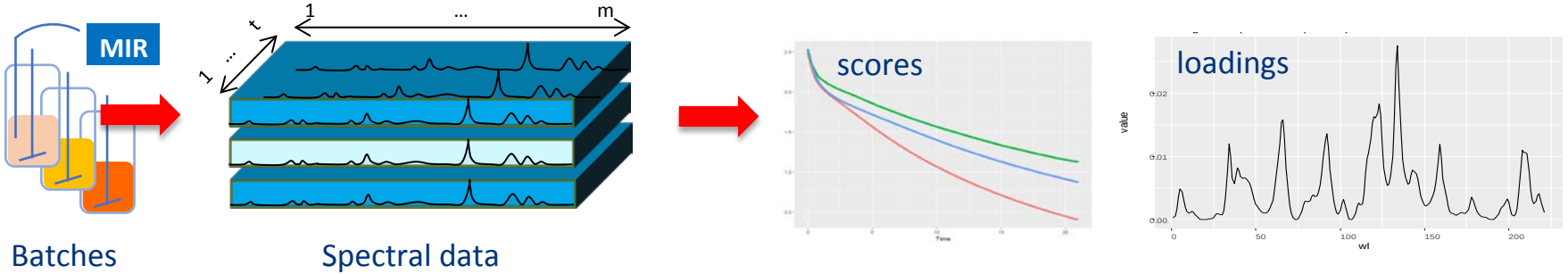
a) single batch



Objective 1: visualization and interpretation

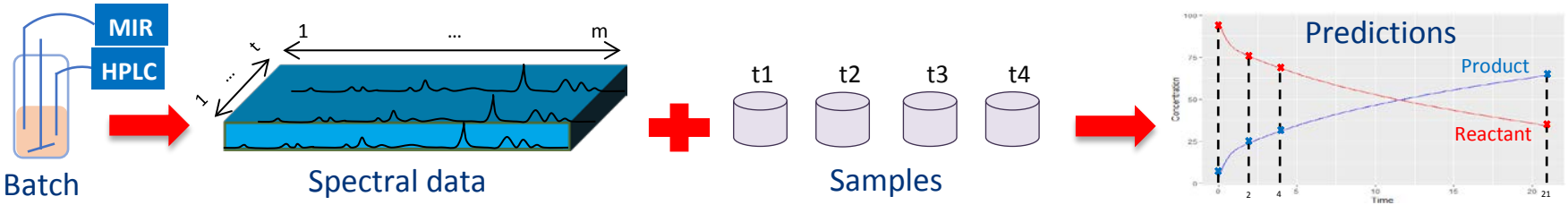
a) single batch

b) multiple batches



Objective 2: prediction of chemical concentrations

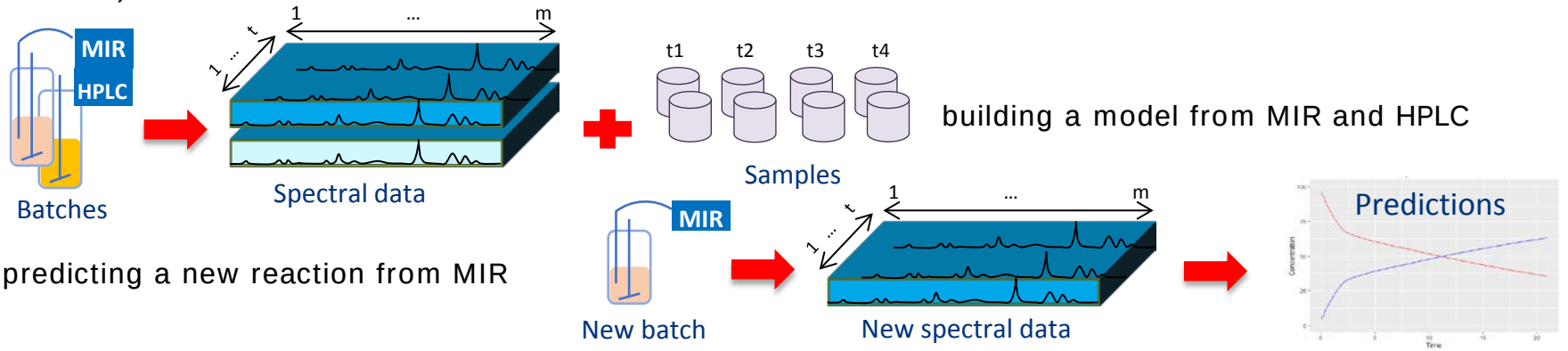
a) single batch
building a model from MIR and HPLC and interpolating



Objective 2: prediction of chemical concentrations

a) single batch
building a model from MIR and HPLC and interpolating

b) for several batches



Methods for dimension-reduction and calibration

Principal component analysis (PCA)

Singular value decomposition of a centered matrix X^c

$$X^c = H \times W'$$

with $X^c \in \mathbb{R}^{t \times m}, H \in \mathbb{R}^{t \times r}, W \in \mathbb{R}^{m \times r}$

Number of dimensions $r = \min(t, m)$

Scores H

Can help to follow reaction components over time

Loadings W

Can help to detect influential spectroscopic variables

Nonnegative matrix factorization (NMF) and Multivariate curve resolution (MCR)

Considering X a non-negative matrix

$X = H \times W' + E$ where H and W are non-negative matrices

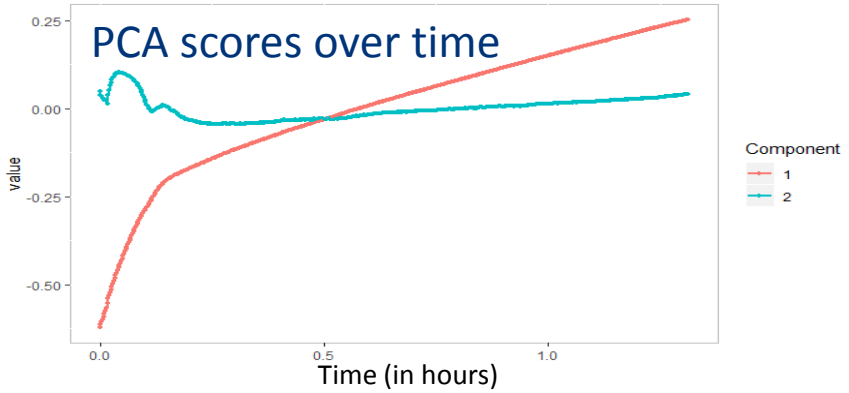
and $X \in R_+^{t \times m}, H \in R_+^{t \times r}, W \in R_+^{m \times r}, E \in R_+^{t \times m}$

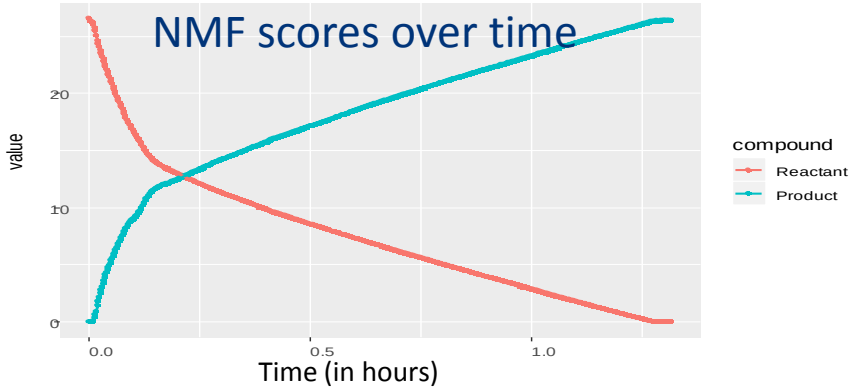
Number of dimensions r chosen upfront

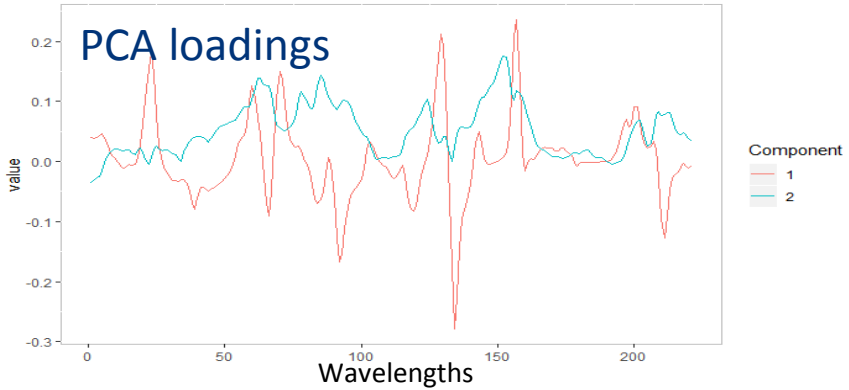
Algorithm

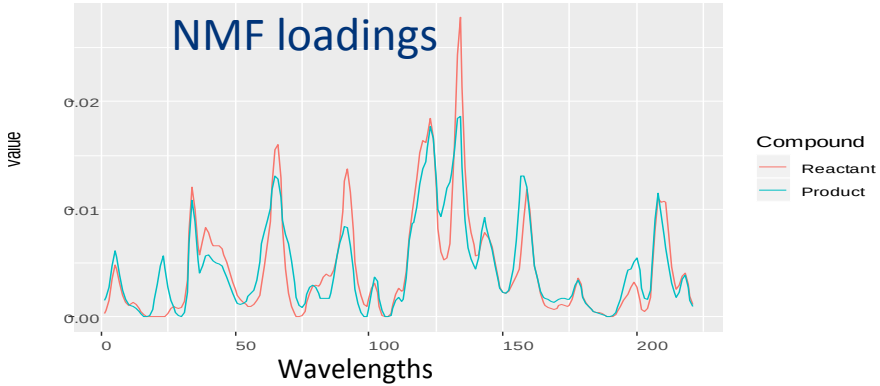
1. Initialize H_0 or W_0
   - Random or prior
   - Pure spectra as prior loadings
2. Optimize H and W alternatively
   - For $k = 0, 1, 2, \ldots$
   - Fix H_k and find W_{k+1} such that $\|X - H_k W'_{k+1}\|_F^2 < \|X - H_k W'_k\|_F^2$
   - Fix W_{k+1} and find H_{k+1} such that $\|X - H_{k+1} W'_{k+1}\|_F^2 < \|X - H_k W'_{k+1}\|_F^2$
3. Stop
   - when $\frac{\|X - HW'\|_F^2}{\|X\|_F^2}$ does not change significantly

Scores and loadings in PCA and NMF









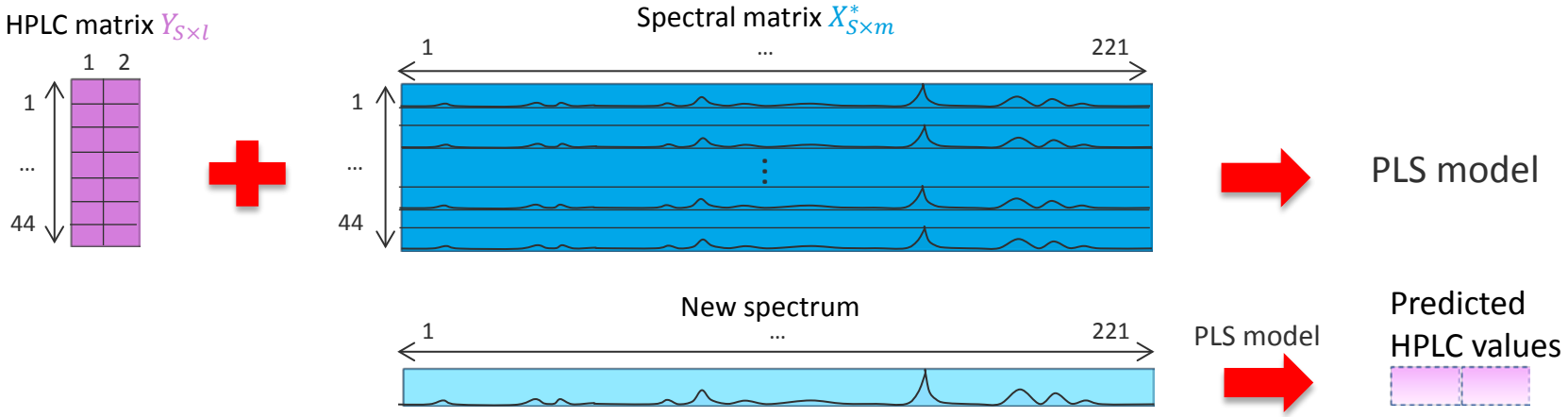
Partial least square regression (PLSR)

Considering $X^*_{S\times m}$ and $Y_{S\times l}$

PLSR is a multiple linear regression method aimed at predicting:

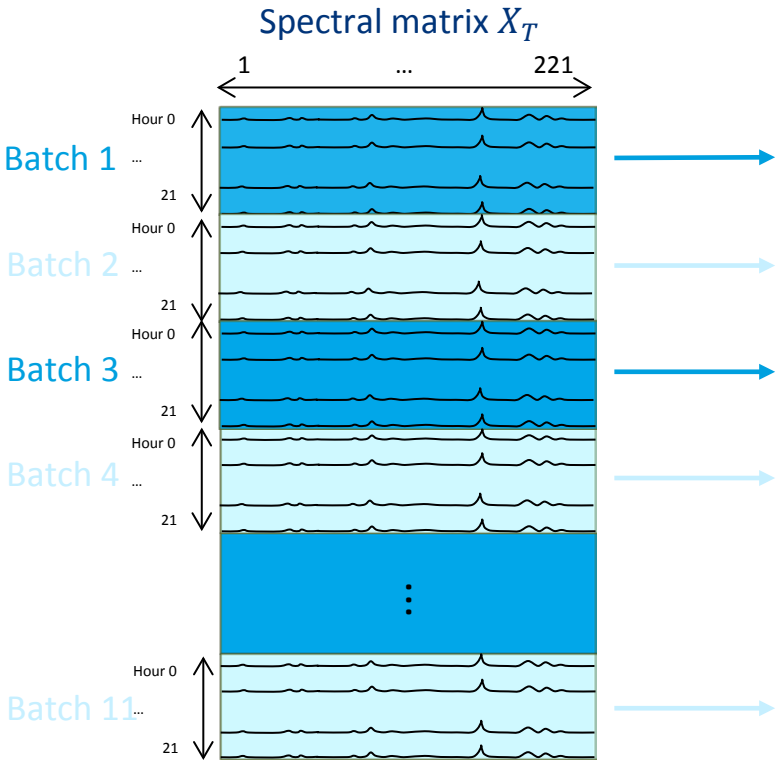
A vector $Y_{S\times 1}$ (PLS1) or matrix $Y_{S\times l}$ (PLS2) from $X^*_{S\times m}$ with $m \gg S$

$$Y_{S\times l} = X^*_{S\times m}\theta_{m\times l} + E_{S\times l}$$



Visualization and interpretation of chemical reactions

Individual models

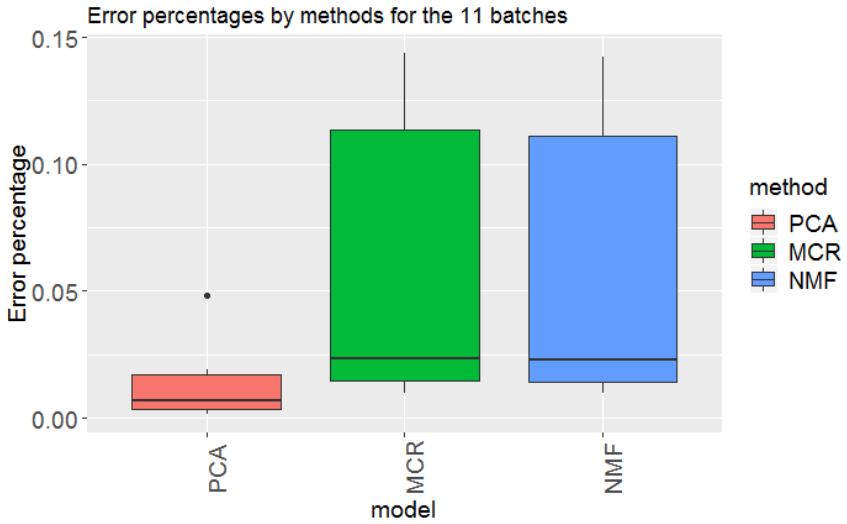


Applying the decomposition on each X_i

$$X_i = H_i \times W_i' + E_i$$

$$\%\ error_i = \frac{\|X_i - H_i \times W_i'\|_F^2}{\|X_i\|_F^2} \times 100$$

Comparison of error percentages



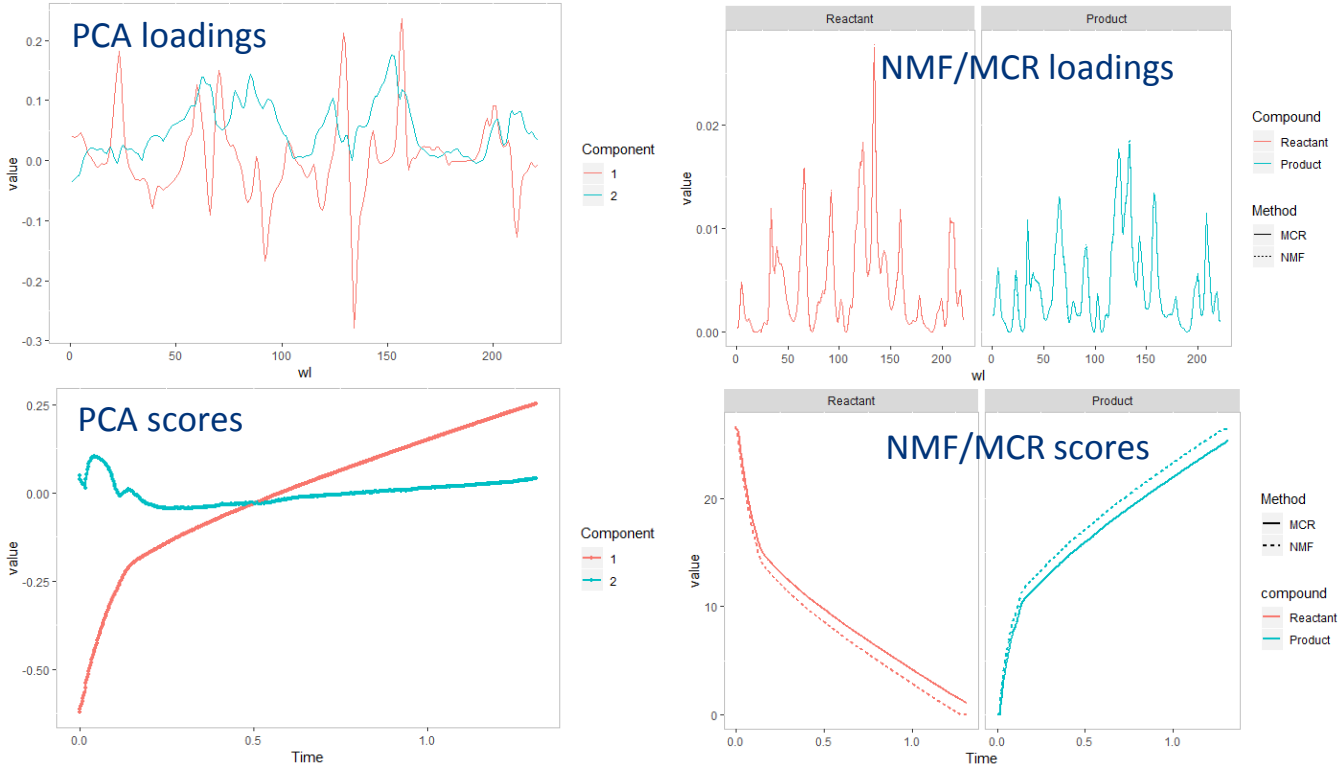
method	Mean error percentage
PCA	0.012%
MCR	0.061%
NMF	0.060%

PCA, MCR and NMF with two components

MCR and NMF have similar errors

PCA has the lowest errors

Comparison among methods for reaction 3



Method	Error percentage
PCA	0.007%
MCR	0.016%
NMF	0.016%

Lower error for PCA

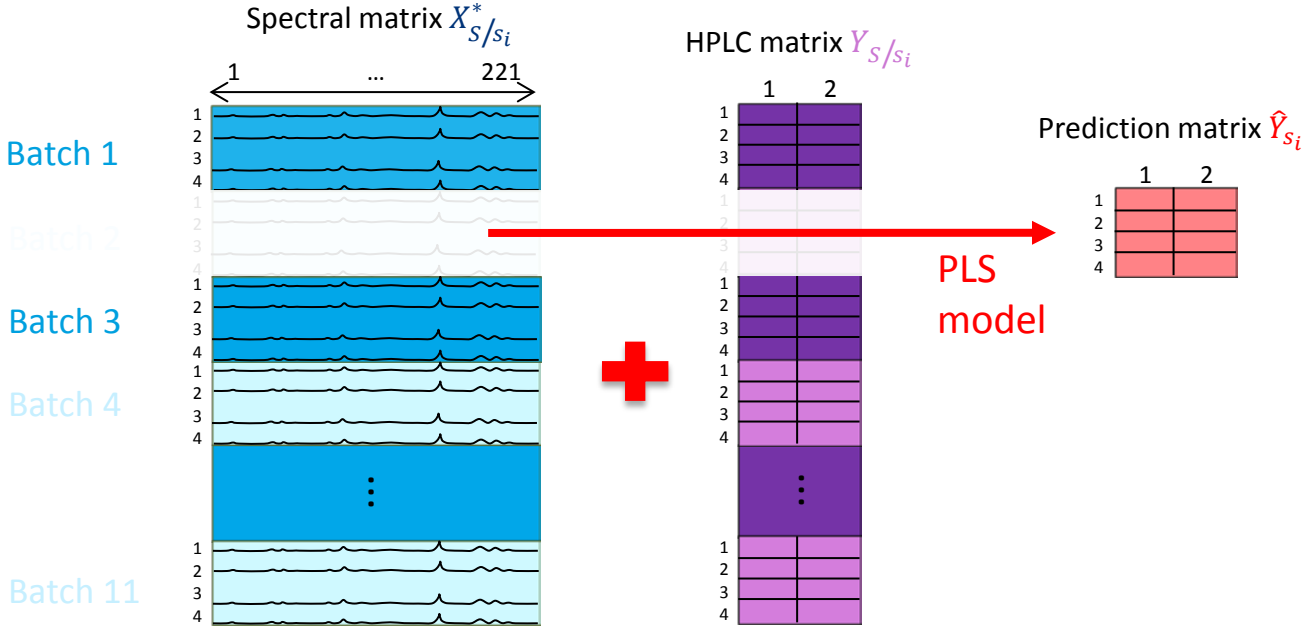
NMF/MCR more interpretable than PCA

Prediction of chemical concentrations

One-step leave one batch out PLS

1. Building a PLS model from X^*_{S/s_i} and Y_{S/s_i}
2. Getting predictions $\hat{Y}_{s_i}$ from $X^*_{s_i}$

Prediction error for a new reaction from previous ones → quality control

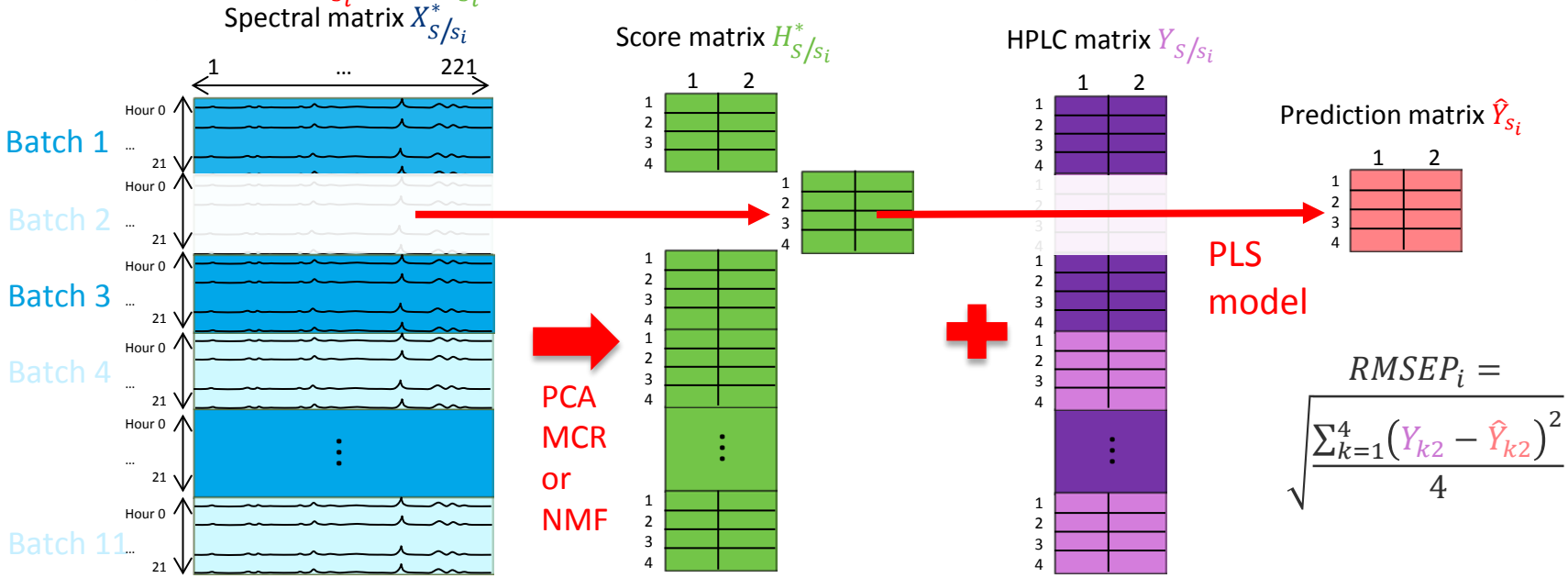


$$RMSEP_i = \sqrt{\frac{\sum_{k=1}^{4}\left(Y_{k2} - \hat{Y}_{k2}\right)^2}{4}}$$

Two-steps leave one batch out PLS

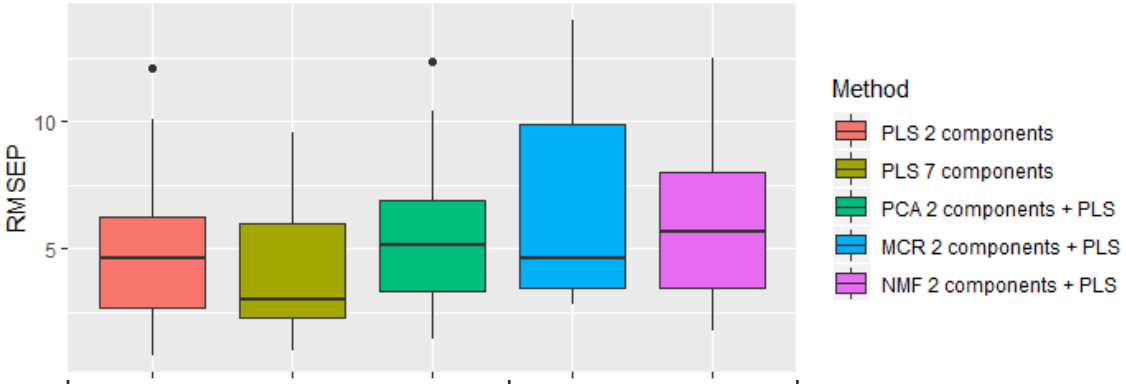
1. Getting scores H^*_{S/s_i} from X_{T/t_i}, then $H^*_{s_i}$ from X_{t_i}
2. Building a PLS model from H^*_{S/s_i} and Y_{S/s_i}
3. Getting predictions $\hat{Y}_{s_i}$ from $\hat{H}^*_{s_i}$

Prediction error for a new reaction from previous ones based on scores → quality control



$$RMSEP_i = \sqrt{\frac{\sum_{k=1}^{4}\left(Y_{k2} - \hat{Y}_{k2}\right)^2}{4}}$$

Comparison of prediction errors for the product



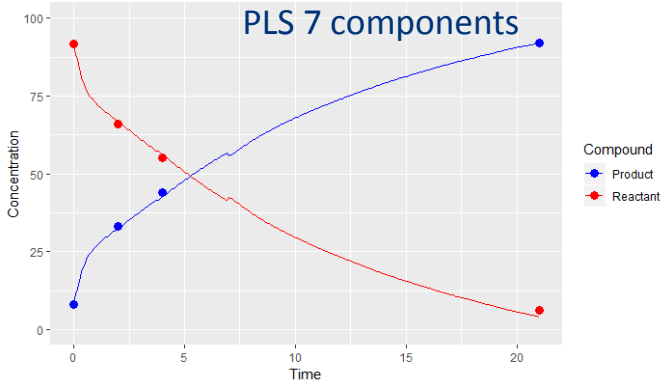
Method	Mean RMSEP
PLS 2 variables	5.10
PLS 7 variables	4.21
PCA 2 variables + PLS	5.71
MCR 2 variables + PLS	7.06
NMF 2 variables + PLS	6.03

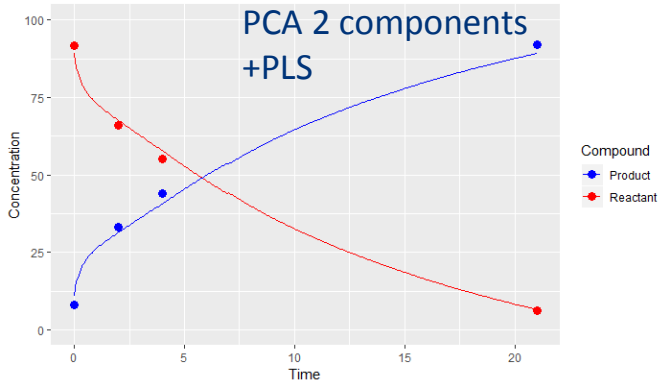
Similar prediction errors from PCA, NMF and MCR scores

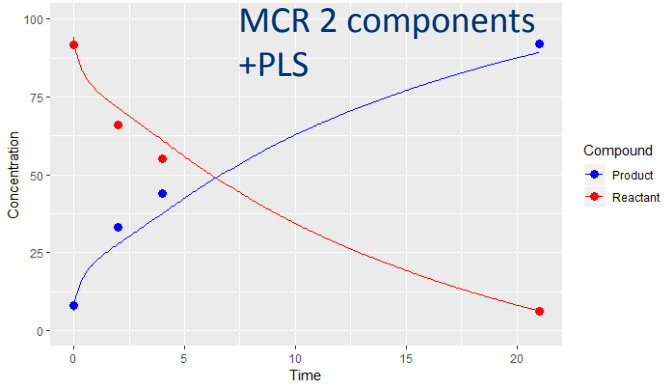
Lowest prediction errors for PLS

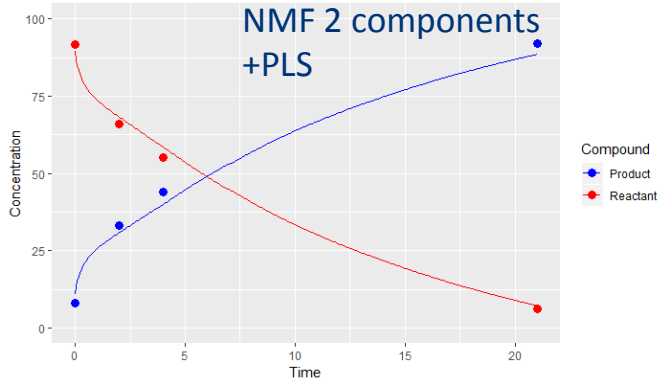
PCA, NMF and MCR scores are informative

Predictions for batch 1









Conclusion and perspectives

Comparison of chemometrics methods

- Better understanding of chemical reactions

Visualization and interpretation of chemical reactions

- NMF and MCR more interpretable than PCA

Prediction of chemical concentrations

- PCA, NMF and MCR similar prediction errors
- PLS remains the reference calibration method

Quality control and real time monitoring

Acknowledgments

Open Analytics:

Adriaan Blommaert

Nicolas Sauwen

Janssen, Manufacturing and Applied statistics:

Tatsiana Khamiakova

Hans Coppenolle

Janssen, PDDS API SM:

Tor Maes

Thank you for your attention!
Any question?

mthiel2@its.jnj.com

04/10/2018

Rhonda Fenwick, *Time is Now I*
Through her art, Rhonda has explored psoriasis, a chronic skin disorder she has lived with since the age of six.

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson


janssen
PHARMACEUTICAL COMPANIES OF
Johnson & Johnson

Back-up slide

Optimizing H_k and W_k

NMF: package hNMF and algorithm *Projected Gradient NMF*

For $k = 0, 1, 2, \ldots$

$W_{k+1} = W_k - \epsilon_W \frac{\partial_f}{\partial_W}$ and set negative elements to 0

$H_{k+1} = H_k - \epsilon_H \frac{\partial_f}{\partial_H}$ and set negative elements to 0

See article “Projected Gradient Methods for Non-negative Matrix Factorization” from Chih-Jen Lin

MCR: package ALS and algorithm *Alternating Least Squares*

For $k = 0, 1, 2, \ldots$

$W_{k+1} = arg\,min_W \|X - H_k W_k'\|_F$ and set negative elements to 0

$H_{k+1} = arg\,min_H \|X - H_k W_{k+1}'\|_F$ and set negative elements to 0

Additional constraints such as unimodality or constant sum of trends can be used
See book “Resolving Spectral Mixtures” from Cyril Ruckebusch