

# Multivariate statistics for PAT data analysis: an application to IR synthesis monitoring

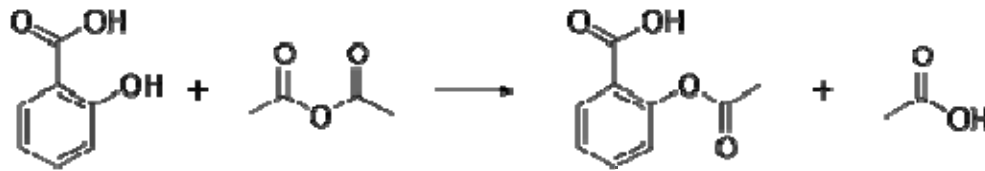
Tatsiana Khamiakova et al.

# Outline

- Introduction to Process analytical technology (PAT)
- IR spectroscopy: challenges is analysis
- Exploratory multivariate analysis
- Exponential model for kinetic rate of reaction
- Discussion

# API synthesis: reaction monitoring

- Reaction for the synthesis of aspirin:



- Goal of the process chemist:
  - ✓ Optimize reaction conditions
  - ✓ Maximize yield
  - ✓ Minimize side products
  - ✓ Maintain robustness of reaction for long period of time (up to 24 h)

<https://en.wikipedia.org/wiki/Aspirin>

# PAT and regulatory agencies

FDA (2004):

*Contains Nonbinding Recommendations*

## **Guidance for Industry<sup>1</sup>**

### **PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance**

This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

# Process Analytical Technology

- PAT is a system to monitor and control the process, e.g. synthesis of API in the lab



‘Offline’ way: take a sample and measure in a separate lab

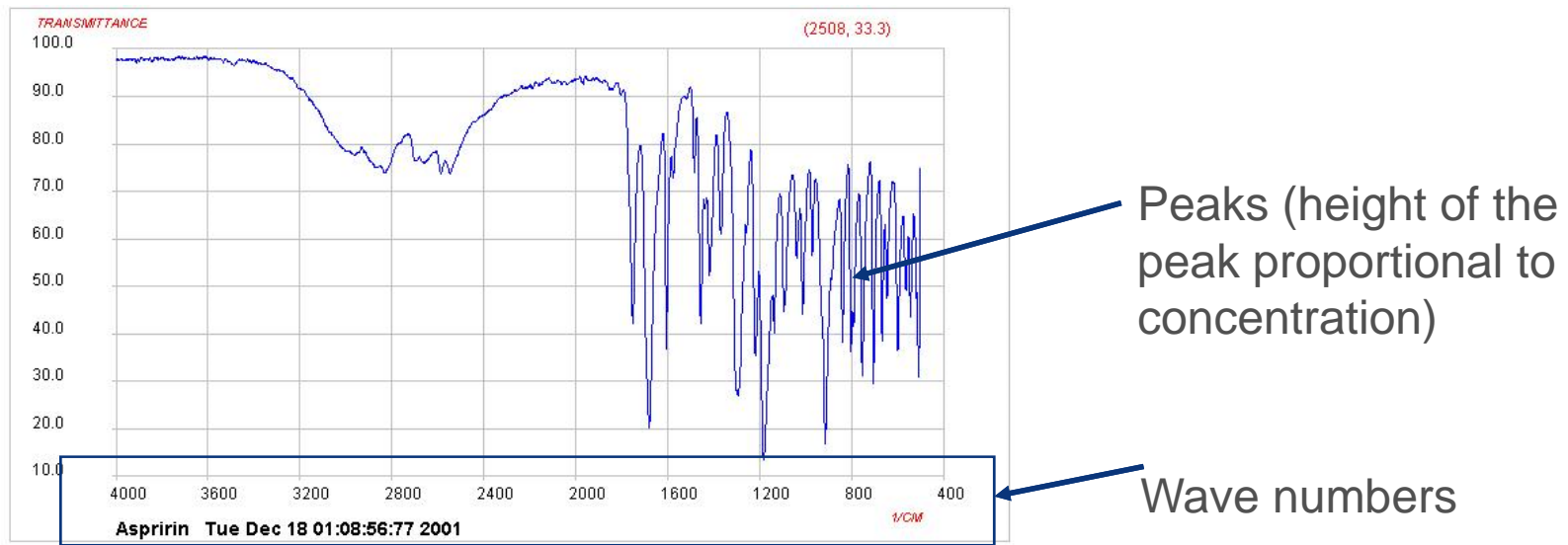


‘Online’ way: measure during reaction

- Advantages of PAT:
  - ✓ Fast
  - ✓ Not laborious
  - ✓ Continuous monitoring

# PAT: (mid)infrared (m-IR) spectroscopy

- m-IR continuously measures relative concentrations of reaction components
- For each measurement a spectrum of values is provided



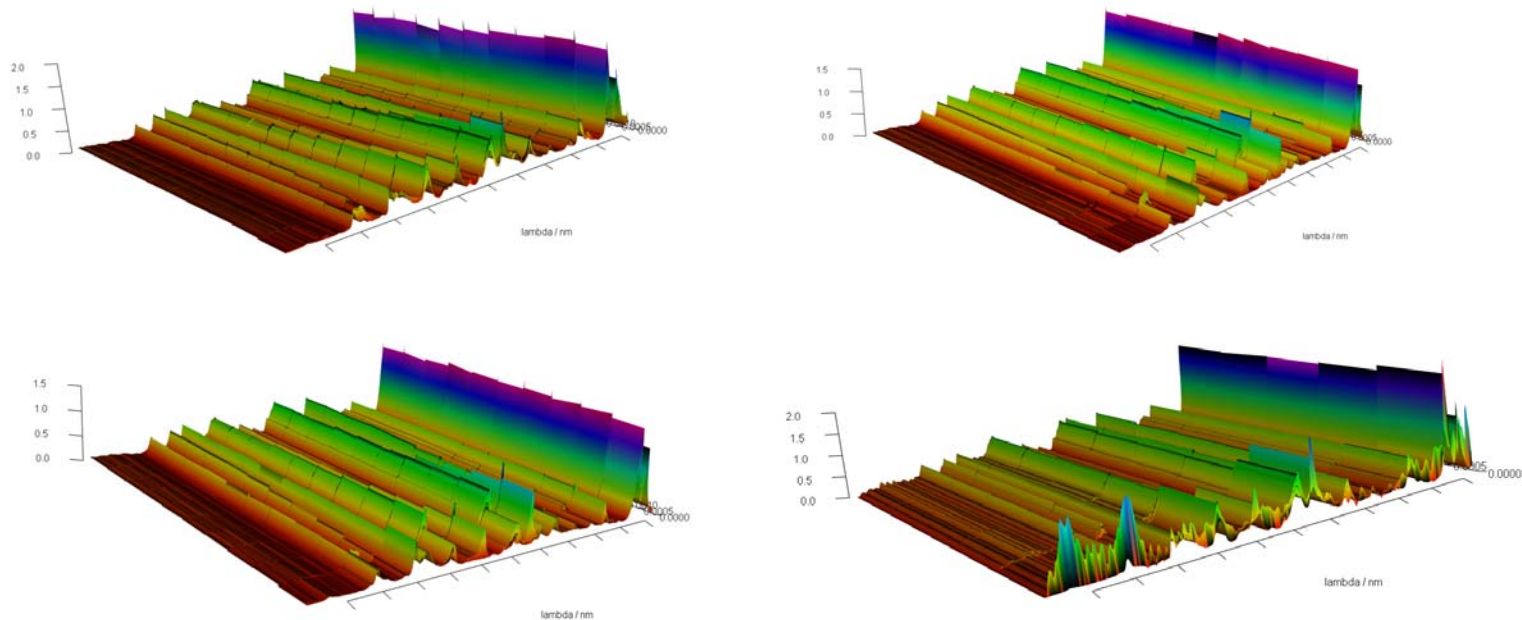
- Spectral signal = series of highly correlated peaks ordered by wave numbers

# Challenges in analysis of spectroscopy data

- Single component may have multiple peaks
- Overlapping peaks
- Complexity of the signal may depend on reaction conditions (concentration of solvents, temperature)
- Need to deconvolve complete spectrum into several components which would (ideally) correspond to the chemical components of a reaction
- Standard techniques in chemometrics (e.g. partial least squares) require offline measurements to calibrate the model

# Data

- Several IR experiments, each containing spectral data for 10 to 20 hours (in total >1000 spectra per experiment)
- Preprocessed by spectralAnalysis R package in-house developed together with Open Analytics (baseline corrected, normalized to the reference peak):





# Exploratory multivariate analyses

## Goals

- investigate the evolution of main reaction components with time to have a quick view on the reaction progress
- investigate the end point detection of a reaction without having offline data

## Methods

- Principal component analysis (PCA)
- Factor analysis for bicluster acquisition (FABIA)
- Non-negative matrix factorization (NMF)

# Methods

- PCA, FABIA, NMF and time series FA: dimensionality reduction (compression) techniques
- General idea: matrix decomposition into  $p$  components

$$Y = \Lambda \times Z + \Psi \text{ or } Y = L \times S,$$

where  $\Lambda_{m \times p}$  and  $L_{m \times p}$  contains loadings (per each point on a spectrum) and  $Z_{p \times n}$  and  $S_{p \times n}$  contains scores of components (per each measured time point)

- The solution is not unique, so different methods apply various restrictions:

**PCA:** orthogonality of components

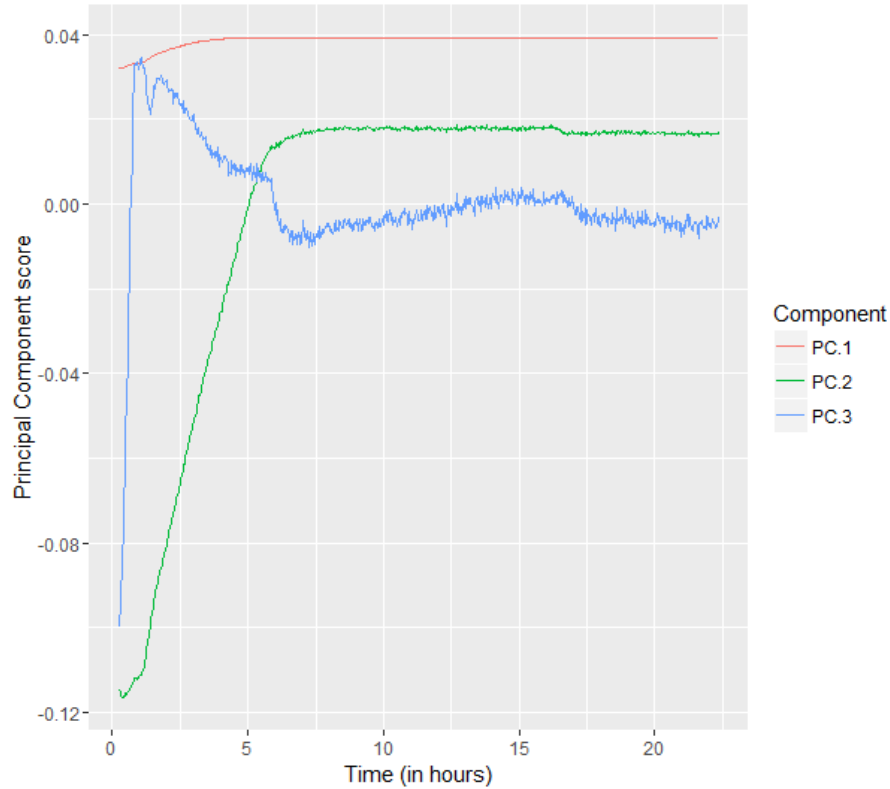
**FABIA:** independence of components and sparseness of loadings

**NMF:** non-negative loadings and scores

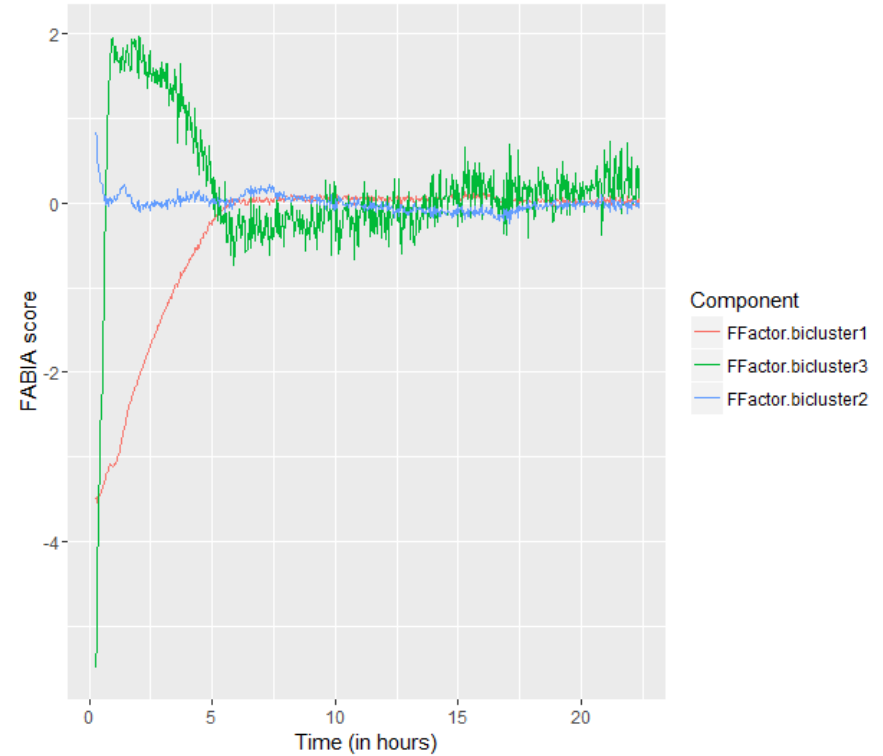
# Challenges in methods application

- **PCA** is sensitive to the irregularities in the measurement process, data may require some pre-treatment
- **FABIA** requires number (upper bound) of the factors to analyze + initialization of algorithm
- **NMF** is sensitive to the number of components and initializations (can use template spectra for the initialization of components)

# PCA and FABIA: illustration on a single run

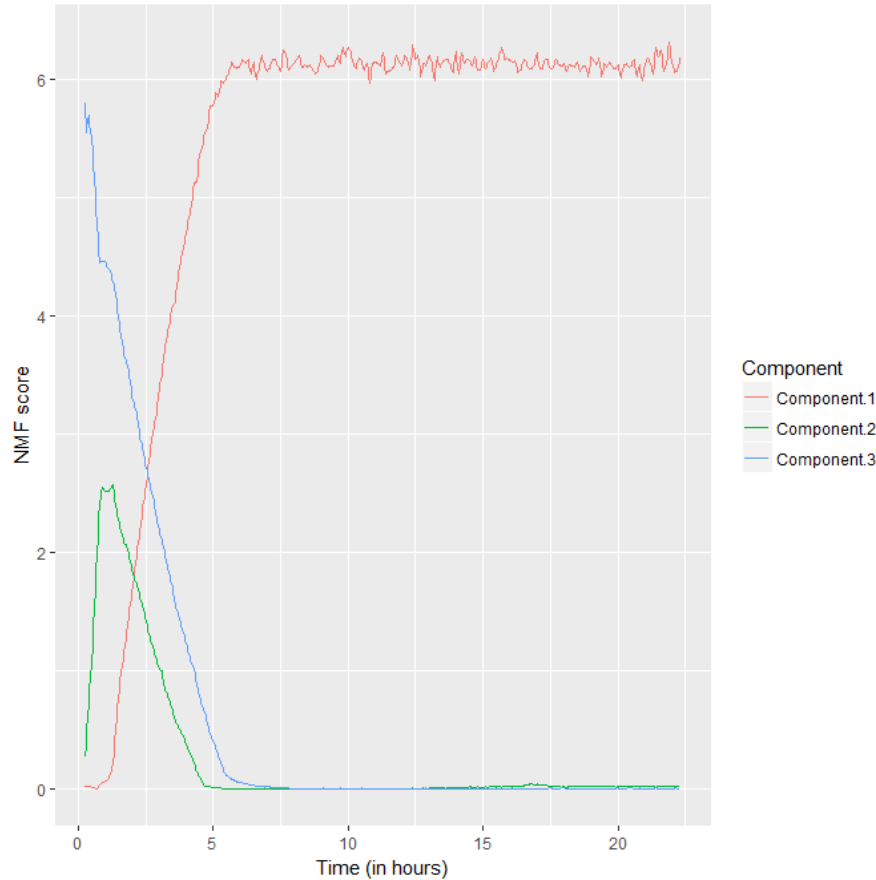


3 principal components from PCA decomposition: the scores

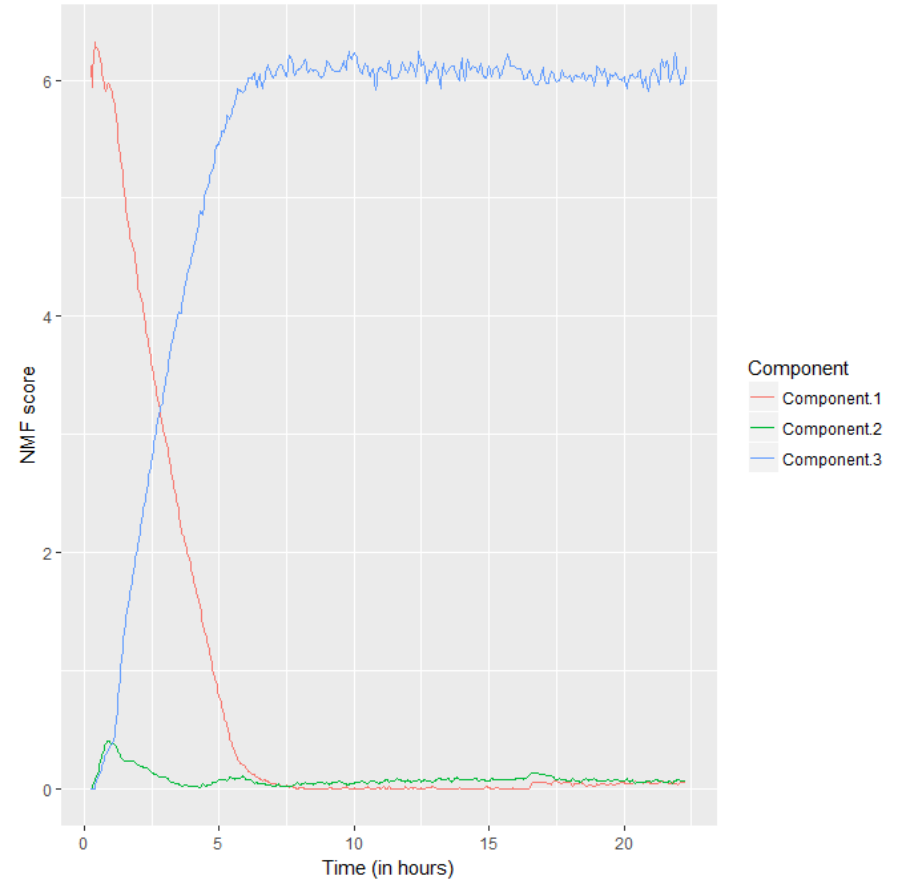


3 factors of FABIA: the scores

# NMF: illustration on a single run



3 components of NMF: the scores based on the initialization with template spectra




3 components of NMF: the scores based on the random initialization

## Discussion: results on a single run

- All methods point out at stabilization of a reaction at approximately 5 hours
- PCA has the most clear reaction trend information in the second component
- FABIA has clear information in the first factor and more noisy trend estimated in the second factor, the third factor looks redundant
- NMF results have the best physico-chemical meaning: non-negative scores and loadings and allow for initialization with template spectra of starting materials and end product
- There are two correlated components (starting material and end product), which makes de composition of this information in different latent factors complex without application of methods like NMF

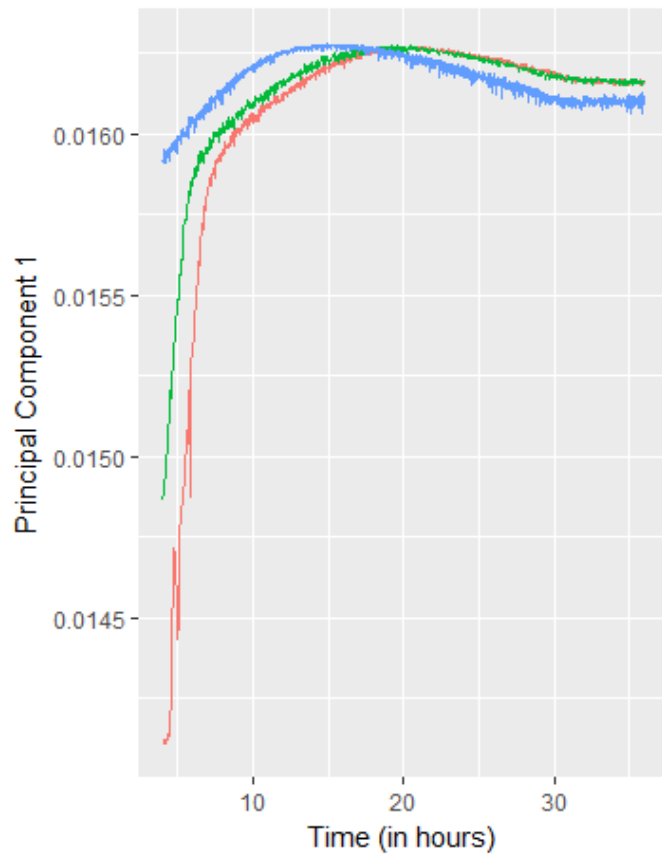
# Methods: analysis of multiple experimental runs

- Often data from several experiments are collected for the same reaction
- Multiple reactions have the same components (i.e. starting materials)  loadings should be common/similar
- To ensure commonality of loadings across experimental conditions, data **from all experiments** are combined in one large matrix:

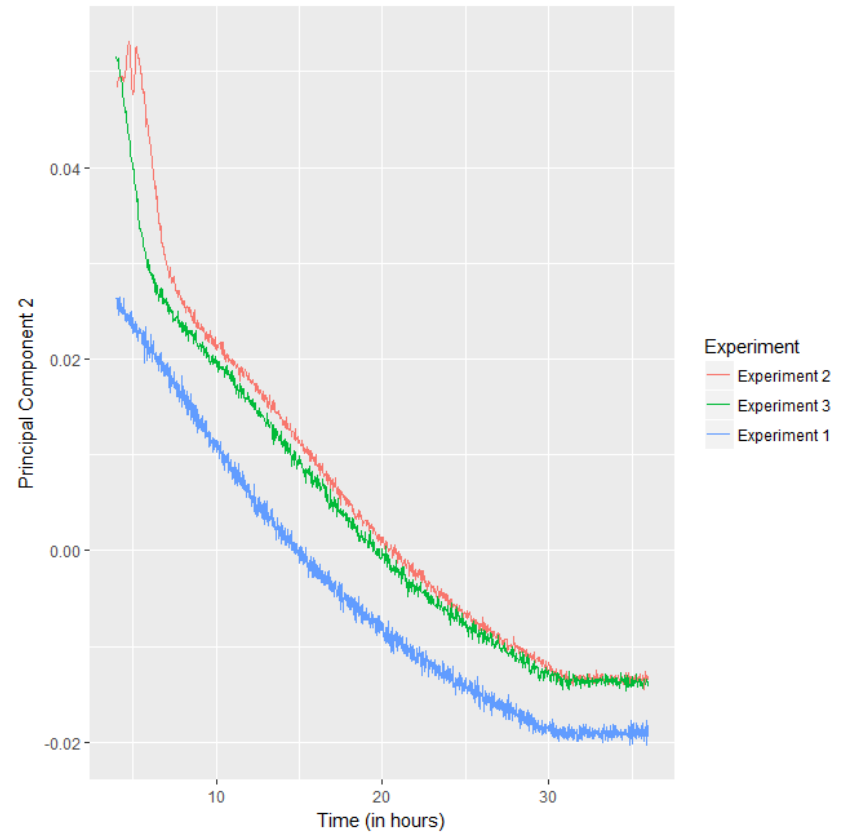
$$Y = (Y_1, Y_2, \dots, Y_k)$$

- It is possible to have separate analyzes per experiment, but loadings will vary and scores interpretation would be more difficult

# PCA on three experimental runs



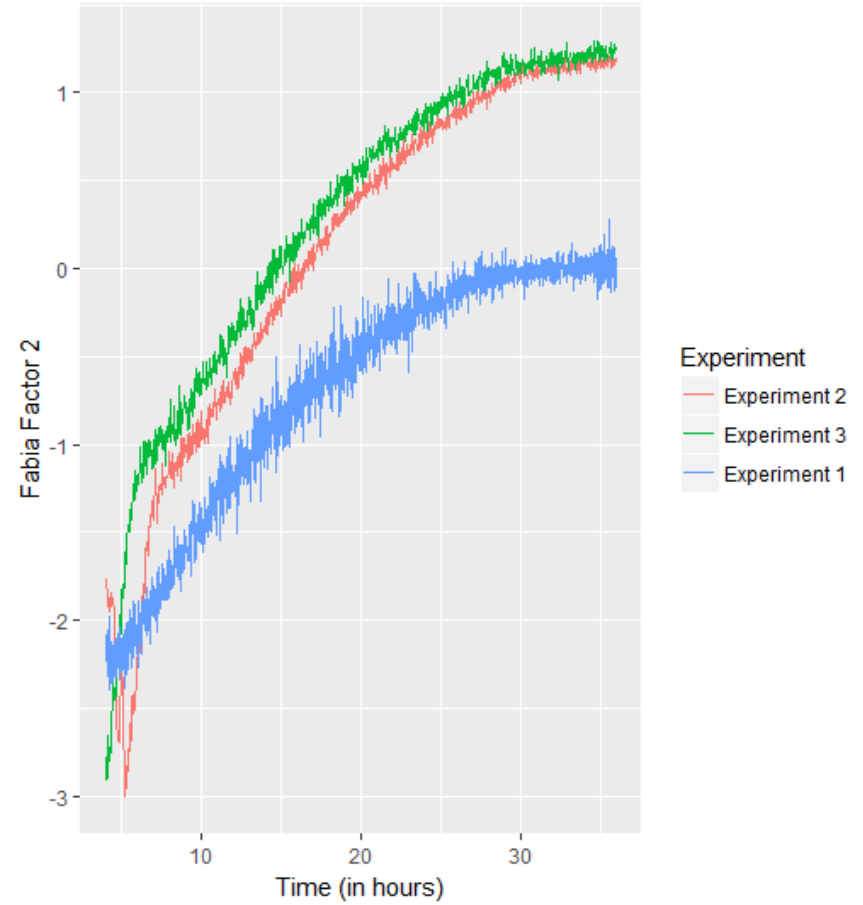
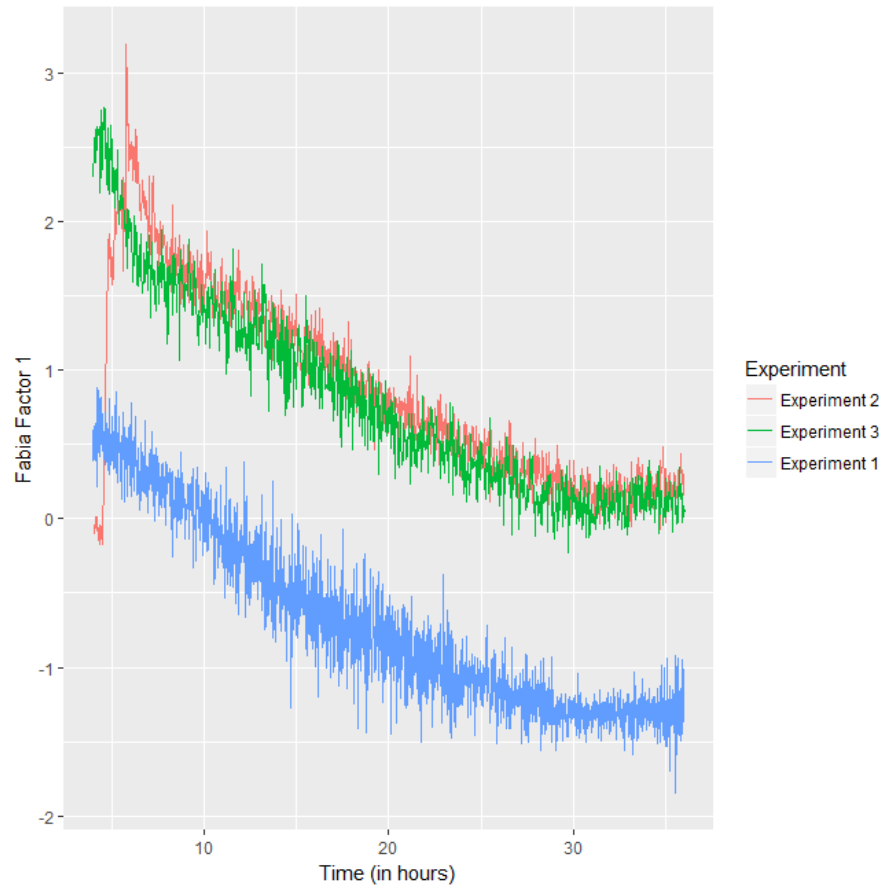
PC1: 97.98%



PC2: 1.56%

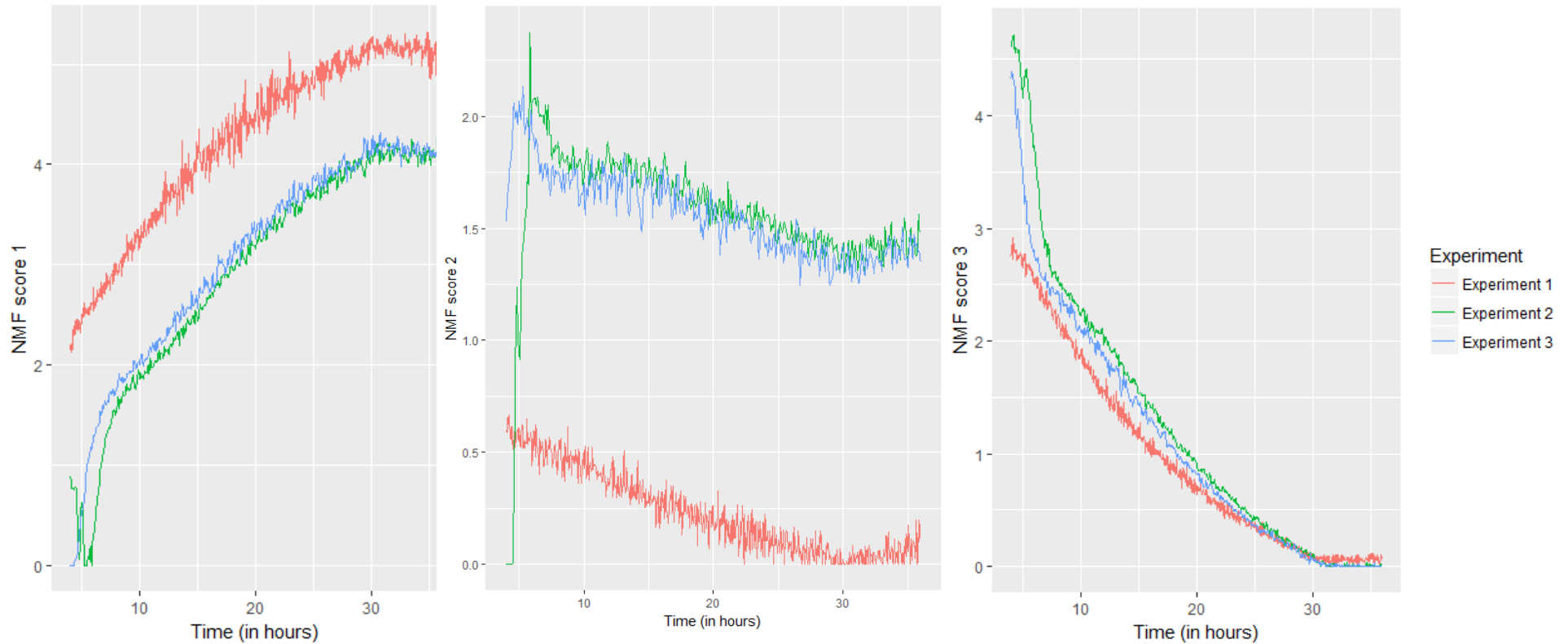


# FABIA on three experimental runs



Two of the three factors have relevant information

# NMF on three experimental runs



If initialization by template spectra is used three components can be identified

## Discussion: analysis of several runs

- Most likely, the first PC contains information on difference between experimental runs (despite normalization), the PC2 contains the reaction progress
- Even though three factors are specified, only two are meaningful in FABIA solution
- NMF provides a decomposition into three different components related to the template spectra of starting materials and end product of the reaction

# Discussion

- Method of dimensionality reduction allow for reaction monitoring without a selection of specific wavelength
- By looking at component-specific trends, the abnormalities in the reaction process can be detected
- These methods do not require offline measurements to decide on the endpoint of a reaction
- PCA is a well-established technique and usually does not require a lot of prior knowledge and experience but is sensitive to variation between experiments
- NMF usually provides more sensible and interpretable results compared to other techniques, but requires more knowledge and experience in order to choose a number of components or initialize the analysis
- NOTE that these methods provide in general qualitative results and are specific for the dataset at hand

## Ongoing research

- Further investigation of how scores can be used for endpoint detection
- Continuous model building for the online monitoring of a trend (so far, successful proof of concept using NMF)
- Calibration transfer problem when shifting from one instrument to the other
- Evaluation of a possibility to make multivariate analysis quantitative (at least to some normalizing constant)

# Conclusions

- IR spectroscopy allows online monitoring of chemical processes
- The generated IR data contains information on various components present in reaction
- Exploring and extracting the signal from the overlapping spectra of components is not a trivial task
- Multivariate techniques can be applied to extract the most relevant components
- The latent structures may or may not have direct chemical interpretation depending on the method at hand
- MVA can be used as a first run analysis to look at the data, check the trends and link individual experiments with experimental conditions

# Acknowledgments

## **Open Analytics:**

Nicolas Sauwen

Adriaan Blommaert

Robin Van Oirbeek

## **Janssen, MTASS-EU:**

Michel Thiel

Helena Geys

## **Janssen, PDMS API SM:**

Tor Maes

Jef Cuypers

Ivan Vervest

Koen De Smet

## References

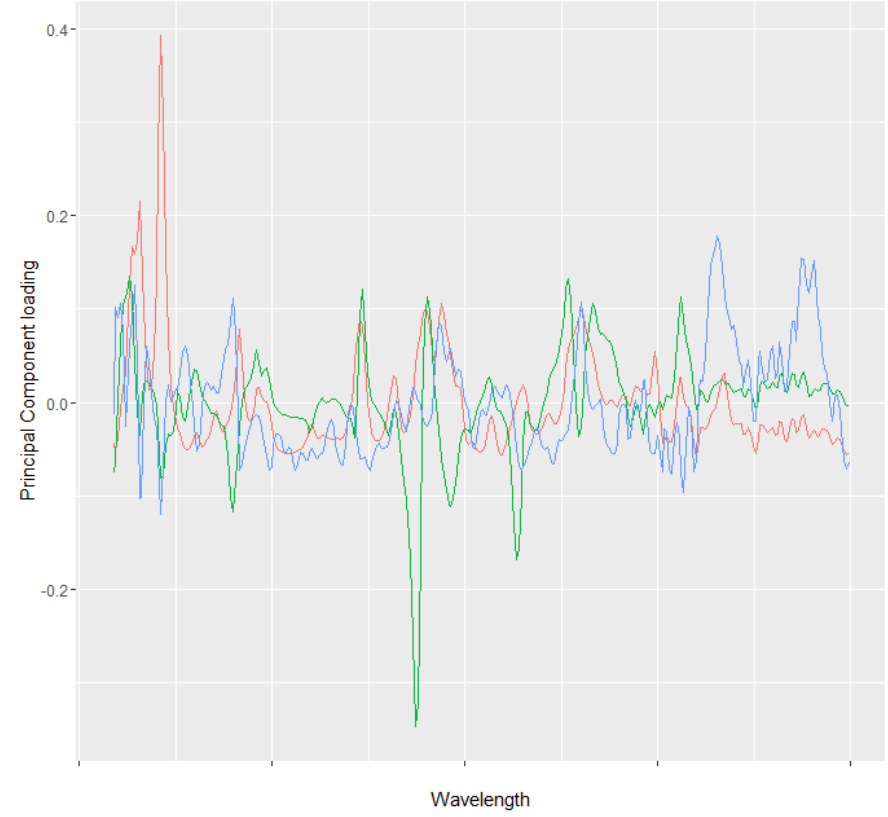
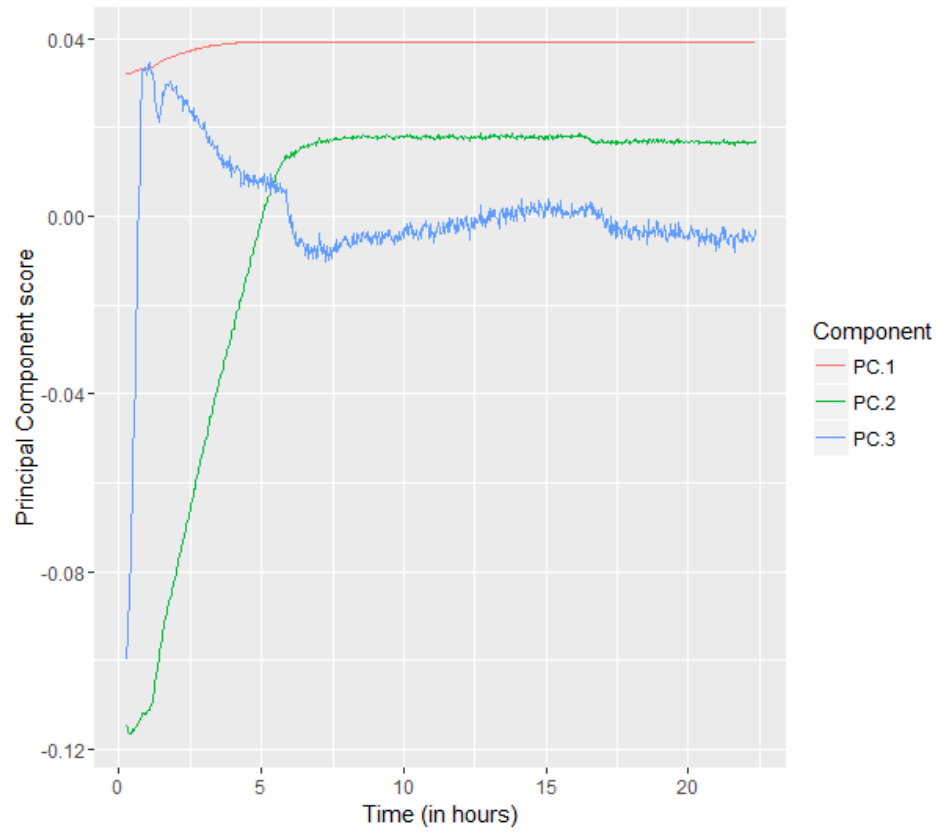
- Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review. (2015) Org. Process Res. Dev. 19, 3–62. [dx.doi.org/10.1021/op500261y](https://doi.org/10.1021/op500261y).
- Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance (2004). FDA guidance: Pharmaceutical CGMPs.
- Hochreiter S., Bodenhofer U., Heusel M., Mayr A., Mitterecker A., Kasim A., Khamiakova T, Van Sanden S, Lin D., Talloen W., Bijmens L., Göhlmann H. H.W., Shkedy Z., Clevert D.-A. (2010) FABIA: Factor Analysis for Bicluster Acquisition. *Bioinformatics*, 26 (12), 1520-1527.
- R package NMF: <https://cran.r-project.org/web/packages/NMF/index.html>



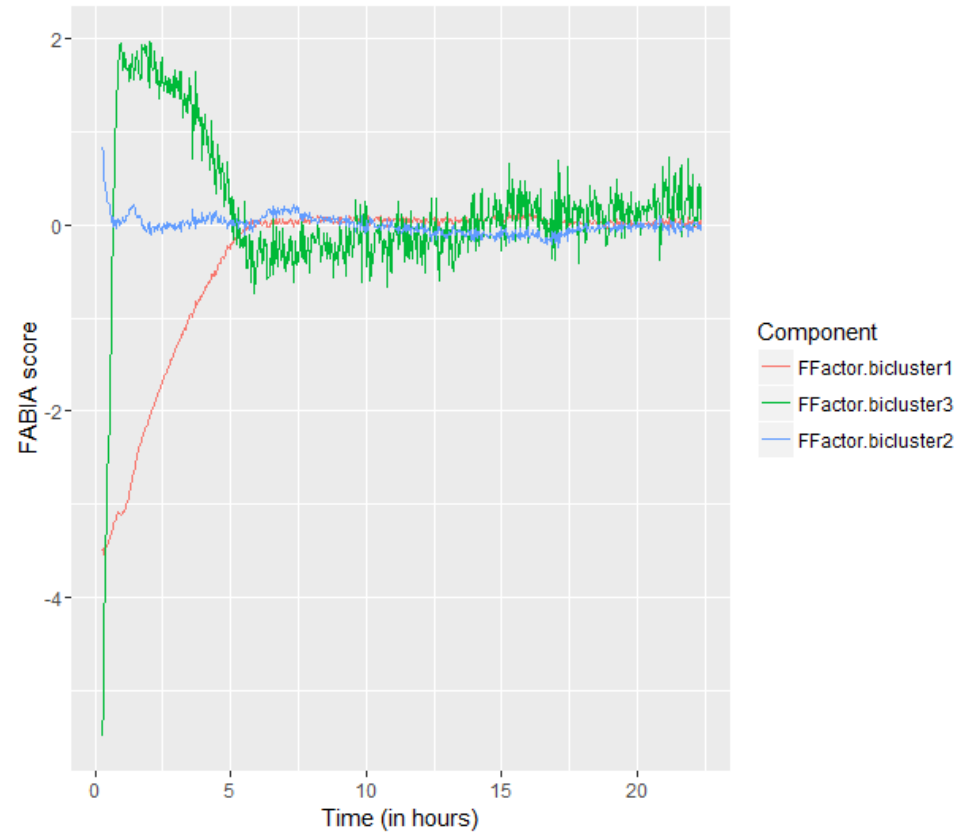
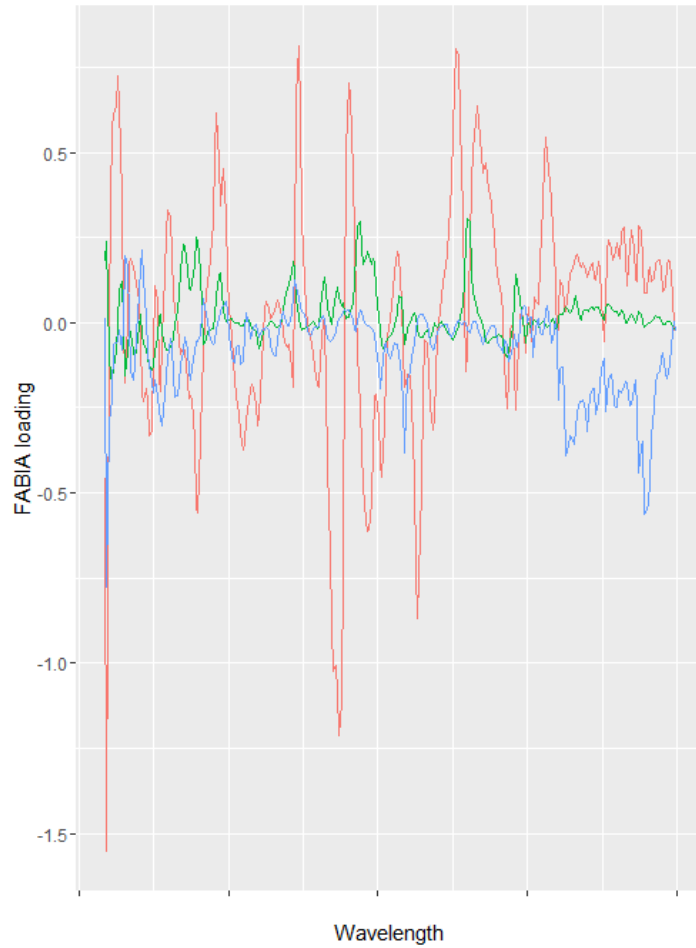
**THANK YOU!**

tkhamiak@its.jnj.com

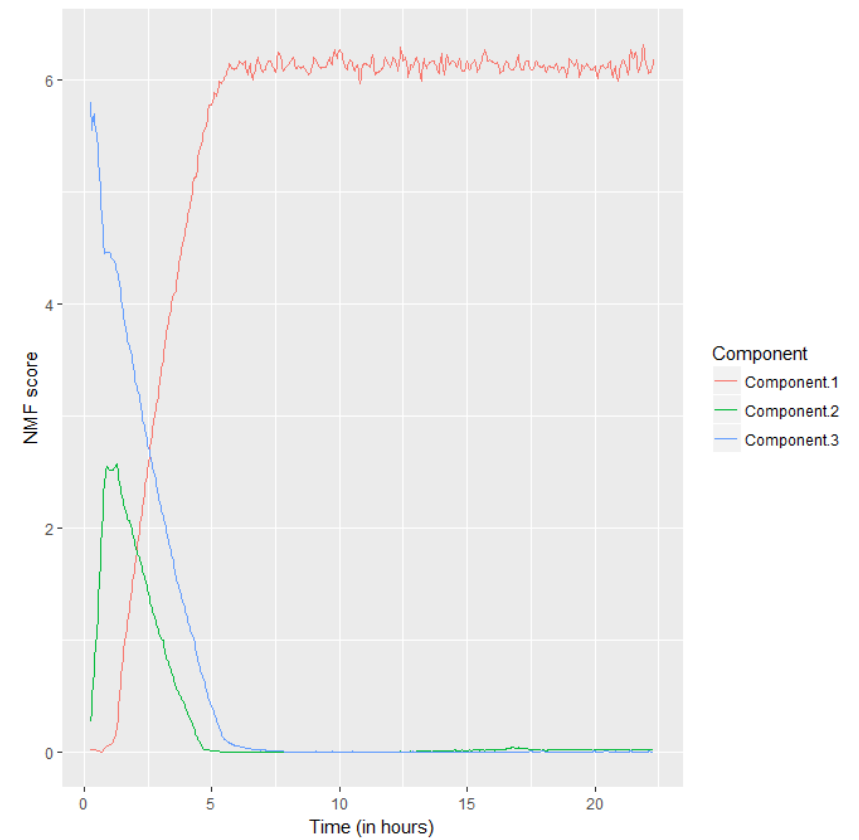
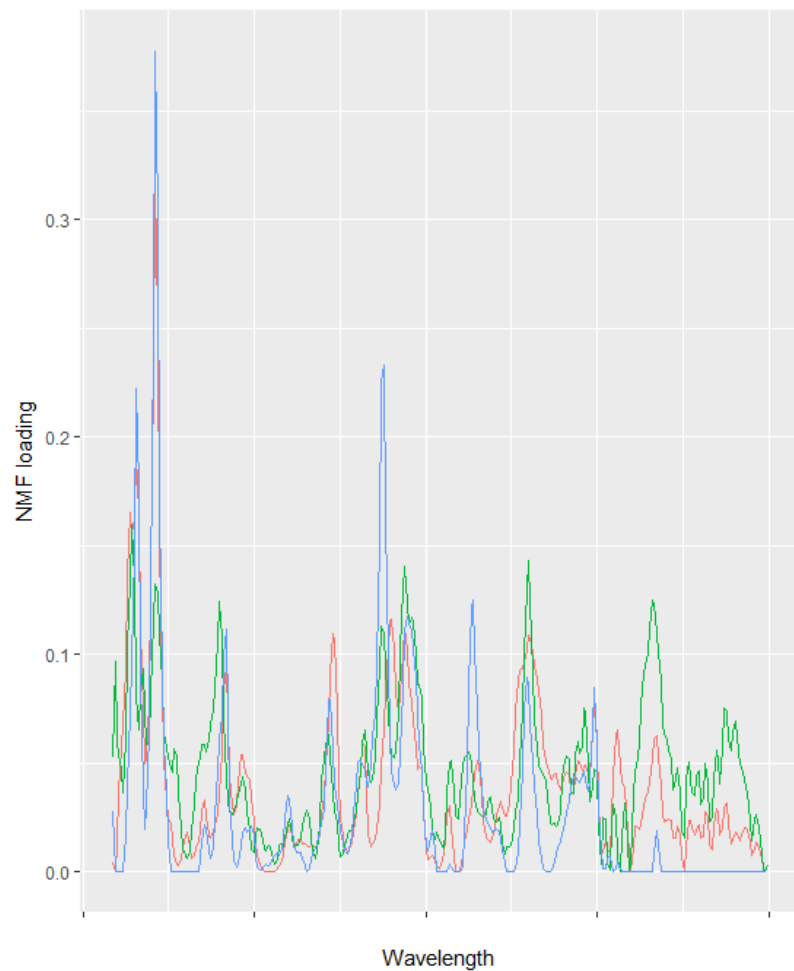
# PCA results



# FABIA results



# NMF results with initialization spectra



# NMF without initialization spectra

