COMPARISON OF CHEMOMETRICS STRATEGIES FOR THE SPECTROSCOPIC MONITORING OF ACTIVE PHARMACEUTICAL INGREDIENTS IN CHEMICAL REACTIONS

Thiel, M., Sauwen, N., Khamiakova, T.,
Maes, T. and B. Govaerts

LIDAM Discussion Paper ISBA/2021/09

Title page

Title: Comparison of chemometrics strategies for the spectroscopic monitoring of active pharmaceutical ingredients in chemical reactions

Authors list:

(Michel Thiel)^{1,4}, Nicolas Sauwen², Tastiana Khamiakova¹, Tor Maes³, Bernadette Govaerts⁴

Affiliations:

¹ Manufacturing and Applied Statistics, Janssen Pharmaceutica, Turnhoutseweg 30, 2340 Beerse, Belgium

² Open Analytics NV, Jupiterstraat 20, 2600 Antwerp, Belgium

³ API Small Molecule Development, Janssen Pharmaceutica, Turnhoutseweg 30, 2340 Beerse, Belgium

⁴ Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Louvain Institute of Data Analysis and Modeling in economics and statistics (LIDAM), UCLouvain, Voie du Roman Pays 20, bte L1.04.01, B-1348 Louvain-la-Neuve, Belgium

Corresponding author:

Michel Thiel^{1,4}, Phone: +3214602575, email: michel.thiel@uclouvain.be

Summary

The Process Analytical Technology (PAT) initiative promoted by the Food and Drug Administration (FDA) encourages pharmaceutical companies to increase the use of new analytical technologies to perform constant monitoring of the critical quality attributes (CQA), allowing a better understanding and a better control of the process.

This paper presents a practical framework based on different dimension-reduction methods as well as calibration methods aimed at following over time chemical experiments organized in batches. To illustrate it, this paper uses pharmaceutical data collected in a research and development context towards industrial production.

This methodological framework aims to reach two objectives.

The first objective is to visualize and interpret in real time, or off-line, the kinetics of chemical reactions using the following dimension-reduction methods: principal component analysis (PCA), non-negative matrix factorization (NMF) and multivariate curve resolution (MCR). The results show that, due to their additional constraints, NMF and MCR allow a better interpretability of chemical reactions than PCA with a comparable quality of fit. Moreover, even though NMF and MCR come from different fields, their algorithms share many similarities and produce close results. The second objective is to predict chemical component concentrations over time. For this second objective, the partial least squares regression (PLSR) is used in a one-step approach and compared with a two-step approach combining multivariate regression with PCA, NMF or MCR. The results show that spectra or scores obtained from unsupervised approaches PCA, NMF or MCR can be used to predict concentrations of the main chemical compounds continuously over all the time of the reaction with a good precision and with a gain of interpretability.

For both objectives, possible model validation indices are also discussed including a leave-one-batch-out approach.

Keywords: Process monitoring, spectroscopy, NMF, MCR, PCA, PLS, PAT

1 Introduction and objectives

The quality attributes of the active pharmaceutical ingredient (API) are critical in the control of pharmaceutical production. Monitoring techniques can be classified as off-line, at-line, and on-line, according to the location of the analytical system in relation to the reactor. Off-line measurements include manual or automatic sampling, transfer to the laboratory, and time delayed analysis in the laboratory. At-line measurements also involve manual or automatic sampling, but the removed sample is analyzed in close proximity to the reactor. For on-line measurements, the reaction is analyzed in real-time by a sensor directly from the reactor (Lourenço et al., 2012).

Off-line techniques such as ultra-pressure liquid chromatography (UPLC) can achieve accurate and precise quantification but their sampling rate is lower and less adequate to enable real-time knowledge of the process performance (Lourenço et al., 2012).

Effective real-time process understanding can be achieved by process analytical technology (PAT). PAT is a system for designing, analyzing, and controlling manufacturing through timely measurements of critical quality and performance attributes with the goal of ensuring final product quality (ICH Guidance for Industry, 2009). In this context, PAT methods such as spectroscopy can be very useful in process monitoring and control, and offer a time-saving alternative (Lapointe-Garant et al., 2008). Spectroscopic methods generate however large datasets with many correlated variables necessitating the use of dimension-reduction methods to extract the meaningful information and correctly monitor the process (De Beer et al., 2011).

This paper presents different chemometrics strategies to monitor chemical reactions under a uniform framework. To illustrate those strategies, multiple chemical reactions were performed in batches to synthesize an active pharmaceutical ingredient. The chemical reactions were monitored on-line as well as off-line in a research and development context as an initial step towards industrial production.

This framework aims to answer two objectives:

The first objective is to visualize and interpret chemical reactions in real time or retrospectively using dimension reduction methods. Those methods can be applied individually to each experiment, globally to compare multiple experiments on a same dimension space, or under a leave-one-batch-out approach to visualize a new experiment using a model built on historical batches.

The second objective is to predict chemical concentrations over time using calibration methods. As for dimension reduction methods, calibration methods can be applied to predict concentration individually to each experiment, globally, or under a leave-one-batch-out approach to predict con-

centrations of a new batch using a calibration set.

The following sections will first describe the dataset used in this article. Secondly, a methodological section will explain the theory behind dimension-reduction methods PCA, NMF and MCR and how to apply them on on-line spectral data organized in batches. The calibration method PLS will then be shortly reviewed and applied to predict off-line chemical concentrations. An original approach in two steps, combining multivariate regression with PCA, NMF and MCR scores will also be presented. For both objectives, model indices are presented to evaluate the quality of the fit and predictions.

Finally, those different methods will be illustrated on the case study.

2 PAT data structure

To motivate the goal of the next sections, it is important to explain the class of data aimed by this article. The data correspond to multiple experiments performed in batches over several hours and monitored by on-line and off-line methods. On-line methods provide continuous but complex high dimensional signals about the process whereas off-line methods provide accurate measurements but their sampling rate is low. The dataset presented in this article should be seen as a typical example of such data set and the framework presented here can be applied on other types of data presenting a similar structure.

2.1 Dataset

The experiments have been performed at Janssen Pharmaceutica in a research and development context to synthesise an active pharmaceutical ingredient. All experiments aimed to produce an active molecule with the same reaction but the process parameters (e.g. temperature, solvent, catalyzer...) were varied in order to understand their impact on the reaction and to select in future steps the optimal production conditions. Eleven experiments were performed in batches monitored over 21 hours on-line by MIR spectroscopy and off-line by UPLC. Since MIR provides spectra with positive intensities, the only pre-processing applied was baseline correction to avoid negative spectral values. MIR and UPLC data can be gathered in two different matrices:

- A spectral matrix $\mathbf{X}_{\mathbf{G}}$ of dimensions $T \times m$ consisting of individual spectral matrices $\mathbf{X}_i$ stacked vertically with $T = \sum_{i=1}^{11} t_i = 6510$ and $m = 221$. Matrix $\mathbf{X}_{\mathbf{G}}$ corresponds to all the 6510 spectra of the 11 experiments put in rows, while the 221 columns correspond to the

spectral variables. The spectral range was between 802 and 1622 nanometers. Table I shows for each experiment the time of monitoring and the number of spectra recorded.

Table I: Summary of spectral measurements per experiment.

Experiment	Time of monitoring (in hours)	Number of spectra
1	21	420
2	21	420
3	21	630
4	21	630
5	21	630
6	21	630
7	21	630
8	21	630
9	21	630
10	21	630
11	21	630

- An UPLC matrix $\mathbf{Y}_{\mathbf{G}}^*$ of dimensions $S \times n$ consisting of individual matrices $\mathbf{Y}_{\mathbf{i}}^*$ stacked vertically where $S = \sum_{i=1}^{11} s_i = 44$, $s_i = 4$ and $n = 2$.

Note then that, for each of the eleven experiments, samples have been taken at only four time points (0, 2, 4 and 21 hours), and, for each sample, the concentrations of three compounds corresponding to the starting material, the final product and a side product were quantified. The side product was present in very low concentration and will then not be used in data analysis of this paper. n will then be considered equal to 2.

Two other matrices $\mathbf{X}_{\mathbf{G}}^*$ of dimensions $S \times m$ and $\mathbf{X}_{\mathbf{i}}^*$ of dimensions $s_i \times m$ can also be respectively defined as the subsets of $\mathbf{X}_{\mathbf{G}}$ and $\mathbf{X}_{\mathbf{i}}$ with the spectra corresponding to off-line sampling times.

The structure of the matrices $\mathbf{X}_{\mathbf{G}}$ and $\mathbf{Y}_{\mathbf{G}}^*$ can be visualized in Figure 1.

Figure 1: Representation of on-line and off-line data.

3 Methods

This section describes first the theory behind dimension-reduction methods PCA, NMF and MCR. It is explained how those methods can be applied on PAT data organized in batches with individual or global models, but also with an original approach called in this article the leave-one-batch-out model. This last approach aims to train a model based on historical batches and apply it on a new one. For each of those approaches, a way to analyze the quality of fit is proposed. The theory behind the calibration method PLS is then described and how this method can be implemented in a one-step approach to predict off-line values from spectral data. A two-step approach combining multivariate regression with PCA, NMF or MCR is also presented, and for both approaches indices to compare the prediction quality are proposed.

3.1 PCA, NMF and MCR dimension-reduction methods

This section summarizes under a common framework the 3 dimension reduction techniques: principal component analysis (PCA), non-negative matrix factorization (NMF) and multivariate curve resolution (MCR) with the perspective of applying them to PAT data.

Let's assume a generic data matrix named $\mathbf{X}$ with dimensions $T \times m$ where T denotes the number of rows and m represents the number of columns. Put differently, $\mathbf{X}$ contains T observations with each observation being described by m variables. This matrix is named with the generic notation $\mathbf{X}$ but could correspond to matrices $\mathbf{X}_G$ or $\mathbf{X}_i$ described in the previous section.

The aim of the different dimension-reduction methods is to decompose the matrix $\mathbf{X}$ into the product of two factor matrices $\mathbf{H}$ of dimensions $T \times R$ and $\mathbf{W}$ of dimensions $m \times R$, providing a low-dimension (number of components R) approximation such that $R \leq \min\{T, m\}$:

$$\mathbf{X} \approx \mathbf{H}\mathbf{W}'$$

The matrix $\mathbf{X}$ can then be expressed as the sum of $\mathbf{H}\mathbf{W}'$ plus the error of the model $\mathbf{E}$ of dimensions $T \times m$:

$$\mathbf{X} = \mathbf{H}\mathbf{W}' + \mathbf{E} = \mathbf{X}_{app} + \mathbf{E}$$

Scores $\mathbf{H}$ and loadings $\mathbf{W}$ matrices represent rows (observations) and columns (variables) into the space of reduced dimensions. In this context, scores will help to follow reaction components over

time while loadings will help to detect influential spectroscopic variables.

The following notations are used throughout all the article to simplify method comparison:

$\mathbf{X}$	$T \times m$	data matrix
$\mathbf{h}_r$	$T \times 1$	$\mathbf{X}$ scores vector r ($r : 1, \dots, R$)
$\mathbf{H}$	$T \times R$	matrix of $\mathbf{X}$ scores vectors
$\mathbf{w}_r$	$m \times 1$	$\mathbf{X}$ loadings vector r ($r : 1, \dots, R$)
$\mathbf{W}$	$m \times R$	matrix of $\mathbf{X}$ loading vectors
$\mathbf{E}$	$T \times m$	Decomposition residual matrix
R	scalar	number of used components

3.1.1 PCA

PCA, for principal component analysis, is one of the most widely known dimension reduction techniques for quantitative data matrices. PCA defines a new coordinate system formed by orthogonal axes called the principal components (PCs) that aim to resume the original variables. Those principal components are new variables constructed as linear combinations of the initial variables. The combinations are chosen such that the principal components are orthogonal to each other and that the first ones contain most of the information available in the data matrix (Varmuza & Filzmoser, 2016).

As seen above, PCA decomposition is defined as:

$$\mathbf{X} = \mathbf{H}\mathbf{W}' + \mathbf{E} = \mathbf{X}_{app} + \mathbf{E}$$

and the scores matrix is calculated as:

$$\mathbf{H} = \mathbf{X}\mathbf{W}'^{-1}$$

Loading vectors are orthogonal to each others and normalized such that $\mathbf{W}'\mathbf{W} = \mathbf{I}_R$ and scores vectors are orthogonal to each others such that $\mathbf{H}'\mathbf{H}$ is diagonal.

It is also important to mention that for interpretability, and prior to PCA decomposition, the matrix $\mathbf{X}$ is generally column-centered by the matrix $\mathbf{M}$ which contains the means of each column of $\mathbf{X}$ such that $\mathbf{X}^c = \mathbf{X} - \mathbf{M}$.

PCA is then applied on $\mathbf{X}^c$: $\mathbf{X}^c = \mathbf{H}^c \mathbf{W}^{c'} + \mathbf{E}^c$. $\mathbf{X}$ is also sometimes standardized when variables have different units. The choice of the number of components to use is done afterwards based on the eigen values, in order to consider enough explained variance.

Several equivalent algorithms are available to perform PCA decomposition such as the singular value decomposition (SVD) or the Non-linear iterative partial least-squares (NIPALS) algorithm (Varmuza & Filzmoser, 2016).

3.1.2 NMF

Non-negative matrix factorization (NMF) was first introduced by Paatero and Tapper in 1994 (Paatero & Tapper, 1994). It became popular after a publication by Lee and Seung in 1999, in which they proposed a simple multiplicative update algorithm for finding a meaningful parts-based representation of facial images (consisting of pixels which have positive intensities) (Lee & Seung, 1999).

NMF is applicable to a non-negative input matrix $\mathbf{X}$, and will produce non-negative factor matrices $\mathbf{H}$ and $\mathbf{W}$. The NMF decomposition is then defined as:

$$\mathbf{X} = \mathbf{H}\mathbf{W}' + \mathbf{E} = \mathbf{X}_{app} + \mathbf{E}$$

with $\mathbf{X} \in \mathbb{R}_+^{T \times m}$, $\mathbf{H} \in \mathbb{R}_+^{T \times R}$, $\mathbf{W} \in \mathbb{R}_+^{m \times R}$, $\mathbf{E} \in \mathbb{R}_+^{T \times m}$ and $R \leq \min(T, m)$ (Lin, 2007).

As such, the NMF model provides physically interpretable factors due to its non-negativity constraints, which is encountered for a wide range of natural signals (Schmidt et al., 2009). The NMF decomposition algorithms require the number of components R as an input. To choose it, it is first of all suggested to rely on existing information about the data or to draw conclusions from the results obtained from other decomposition methods such as SVD (Boutsidis & Gallopoulos, 2008).

A similarity measure $D(\mathbf{X} || \mathbf{H}\mathbf{W}')$ is required in order to quantify the difference between $\mathbf{X}$ and its approximation $\mathbf{H}\mathbf{W}'$. This similarity measure also serves as a cost function for solving the optimization problem. The most commonly used cost functions are based either on Squared Euclidean Distance (SED) or on Generalized Kullback-Leibler Divergence (GKLD).

Squared Euclidean Distance:

$$D_{SED}(\mathbf{X}||\mathbf{H}\mathbf{W}') = \sum_{ij} (X_{ij} - [HW']_{ij})^2 = \|\mathbf{X} - \mathbf{H}\mathbf{W}'\|_F^2$$

where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix.

Generalized Kulback-Leibler Divergence:

$$D_{GKLD}(\mathbf{X}||\mathbf{H}\mathbf{W}') = \sum_{ij} (X_{ij} \log \frac{X_{ij}}{[HW']_{ij}} - X_{ij} + [HW']_{ij})$$

As most scientific articles about NMF are based on the squared euclidean distance, this is the cost function that will be developed below.

Although many NMF algorithms exist in the literature, they can be divided into three major categories: gradient descent algorithms, multiplicative update algorithms, and alternating least squares algorithms (Berry et al., 2007). It can be mentioned beforehand that none of these algorithms is expected to find a global minimum for the cost function since the cost function (SED or GKLD) is only convex in either $\mathbf{H}$ or $\mathbf{W}$, but not in the two of them together. Those 3 categories of algorithms use the Frobenius norm as the cost function to be minimized.

More details on those 3 categories can be found in the literature but basically the idea is to fix alternatively one of the factor matrices while optimizing the other one in such a way that the objective function $\|\mathbf{X} - \mathbf{H}_k \mathbf{W}'_k\|_F^2$ is reduced at each iteration k (Gillis, 2014; Lin, 2007).

The first step is the choice of the number of components R . In this context, the number of components R can be chosen to correspond to the number of chemical compounds reacting in the mixture. The idea is that the score matrix $\mathbf{H}$ is proportional to the concentration profiles and that the loading matrix $\mathbf{W}$ corresponds to the spectral profiles. Matrices $\mathbf{H}$ and $\mathbf{W}$ can then be initialized as dense matrices of random numbers between 0 and 1 or by providing prior information such as the expected concentration profiles or the pure component spectra. An example of gradient descent algorithm is as follows:

1. Fix the number of components R .
2. Initialize matrices $\mathbf{H}_0$ and $\mathbf{W}_0$ to non-negative matrices as random or with prior information.
3. Optimize $\mathbf{H}_k$ and $\mathbf{W}_k$ alternately for $k = 0, 1, 2, \dots$

Keep $\mathbf{W}_k$ fixed and find $\mathbf{H}_{k+1}$ such that $\|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}'_k\|_F^2 < \|\mathbf{X} - \mathbf{H}_k \mathbf{W}'_k\|_F^2$

where $\mathbf{H}_{k+1} = \text{element max} \{0, \mathbf{H}_k - \alpha_k \nabla_{\mathbf{H}} \|\mathbf{X} - \mathbf{H}_k \mathbf{W}'_k\|_F^2\}$.

Keep $\mathbf{H}_{k+1}$ fixed and find $\mathbf{W}_{k+1}$ such that $\|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}'_{k+1}\|_F^2 < \|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}'_k\|_F^2$
where $\mathbf{W}_{k+1} = \text{element max} \{0, \mathbf{W}_k - \alpha_k \nabla_{\mathbf{W}} \|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}'_k\|_F^2\}$.

4. Stop iteration when the convergence criterion is met.

α_k is the step-size parameter and influences strongly the convergence of gradient descent method. A wrong choice has consequences ranging from a very huge number of iterations to divergence. As the subtraction operations used in each iteration can lead to a violation of the non-negative property of $\mathbf{H}$ and $\mathbf{W}$, negative elements of those matrices are replaced by 0 at each iteration.

All NMF algorithms are iterative and they are sensitive to the initialization of $\mathbf{H}$ and $\mathbf{W}$ (Langville et al., 2014). Nearly all NMF algorithms use simple random initialization, i.e., $\mathbf{H}$ and $\mathbf{W}$ are initialized as dense matrices of random numbers between 0 and 1 (Langville et al., 2014). Those random initializations are repeated a certain number of time and the solution with the lowest cost function is selected. But another possibility is to provide a prior $\mathbf{H}_0$ or $\mathbf{W}_0$ based on the knowledge of the data (e.g. in PAT). Other types of initialization also exist and are compared in the article of Langville et al., 2014.

There are also several approaches for the stopping criterion of NMF algorithms. Common approaches consist to fix a maximum number of iterations or a time limit to ensure termination of the algorithm (Gillis, 2014).

But other options are available (Cichocki et al., 2009; Erichson et al., 2016; Gillis, 2014): terminate the algorithm if the value of the cost function is smaller than some stopping condition ε :

$$\|\mathbf{X} - \mathbf{H}_k \mathbf{W}'_k\|_F^2 < \varepsilon$$

or if there is little or no improvement between successive iterations in the minimization of the cost function, for example:

$$\|\mathbf{H}_k \mathbf{W}'_k - \mathbf{H}_{k+1} \mathbf{W}'_{k+1}\|_F^2 < \varepsilon$$

3.1.3 MCR

MCR, for multivariate curve resolution, is the generic name of a family of methods meant to solve chemical mixture problems such as spectroscopic measurements and were first developed in the field of chemometrics (Ruckebusch, 2016).

The approach can be non-iterative and include methods such as Orthogonal Projection Resolution (OPR) and Window Factor Analysis (WFA), or the approach can be iterative and include methods like Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS). In practice, iterative methods are used more frequently and in this paper only MCR-ALS, which is one of the oldest, most well-known and widely used algorithm, will be analyzed in detail.

The MCR decomposition is defined as:

$$\mathbf{X} = \mathbf{H}\mathbf{W}' + \mathbf{E} = \mathbf{X}_{app} + \mathbf{E}$$

with $\mathbf{X} \in \mathbb{R}_+^{T \times m}$, $\mathbf{H} \in \mathbb{R}_+^{T \times R}$, $\mathbf{W} \in \mathbb{R}_+^{m \times R}$, $\mathbf{E} \in \mathbb{R}_+^{T \times m}$ and $R \leq \min(T, m)$

MCR shares then the same objective as NMF, decomposing a nonnegative matrix $\mathbf{X}$ into factor matrices $\mathbf{H}$ and $\mathbf{W}$ and update them under additional constraints so that the objective function $\|\mathbf{X} - \mathbf{H}_k \mathbf{W}_k'\|_F^2$ is reduced at each iteration k . The MCR-ALS algorithm can then be written as:

1. Fix the number of components R .
2. Initialize matrices $\mathbf{H}_0$ and $\mathbf{W}_0$ to non-negative matrices as random or with prior information.
3. Optimize $\mathbf{H}_k$ and $\mathbf{W}_k$ alternately for $k = 0, 1, 2, \dots$.

Fix $\mathbf{W}_k$ and find $\mathbf{H}_{k+1}$ such that $\|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}_k'\|_F^2 < \|\mathbf{X} - \mathbf{H}_k \mathbf{W}_k'\|_F^2$ under constraints where $\mathbf{H}_{k+1} = \text{element max} \{0, \mathbf{X} \mathbf{W}_k (\mathbf{W}_k' \mathbf{W}_k)^{-1}\}$.

Fix $\mathbf{H}_{k+1}$ and find $\mathbf{W}_{k+1}$ such that $\|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}_{k+1}'\|_F^2 < \|\mathbf{X} - \mathbf{H}_{k+1} \mathbf{W}_k'\|_F^2$ under constraints where $\mathbf{W}_{k+1} = \text{element max} \{0, \mathbf{X}' \mathbf{H}_{k+1} (\mathbf{H}_{k+1}' \mathbf{H}_{k+1})^{-1}\}$.
4. Stop iterations when convergence criterion is met.

To ensure non-negativity constraint, the algorithm uses least squares to alternatively optimize $\mathbf{H}$ and $\mathbf{W}$, and negative values are replaced by 0. The optimization is usually stopped when the difference of model fit between consecutive iterations does not improve significantly (e.g., a difference of less than 0.1% among the lack of fit (% LOF) between consecutive iterations, or the percentage of variance explained (% var. expl.)). Other possibilities are using a maximum number of iterations or criteria based on the comparison of shapes of the optimized profiles among iterations (Ruckebusch, 2016).

The lack of fit criteria can be defined as: $\%LOF = 100 \frac{\|\mathbf{E}\|_F}{\|\mathbf{X}\|_F}$

and the percentage of variance explained as: $\%var.expl. = 100(1 - \frac{\|\mathbf{E}\|_F^2}{\|\mathbf{X}\|_F^2})$

Since MCR was developed in the field of chemometrics, this method proposes additional constraints than component non-negativity to adapt to physical constraints encountered to chemical mixtures. Those constraints are for example unimodality, which allows the presence of a single maximum per concentration profile, or closure, which adapts the total sum of the concentration of compounds to a constant value during all the reaction. Those constraints also help to reduce or suppress the ambiguity of solutions encountered with MCR (and NMF). The ambiguity phenomenon can be described such that different combinations of concentration profiles and spectra can describe equally well, in terms of model fit, the original data set (Tauler & Maeder, 2009; Tauler et al., 1995).

3.1.4 Summary of methods PCA, NMF and MCR

Table II briefly summarizes the principal characteristics of methods PCA, NMF and MCR.

Table II: Summary of methods PCA, NMF and MCR.

	PCA	NMF	MCR
Expression	$\mathbf{X}^c = \mathbf{H}^c \mathbf{W}^{c'} + \mathbf{E}^c$	$\mathbf{X} = \mathbf{H} \mathbf{W}' + \mathbf{E}$	$\mathbf{X} = \mathbf{H} \mathbf{W}' + \mathbf{E}$
Objective	Finding new axes (principal components) maximizing the variance of the data	Matrix decomposition into two non-negative matrices	Finding the components of a mixture
Constraints	Orthogonality	Non-negativity of all elements in $\mathbf{H}$ and $\mathbf{W}$	User-defined (e.g. non-negativity, unimodality, closure etc.)
Examples of algorithms	SVD, NIPALS	Multiplicative update algorithms, gradient descent, ALS	MCR-ALS
Prior information	No	Yes	Yes
Main limitations	Interpretation	Ambiguity of the solutions, results influenced by parameter and initialization	Ambiguity of the solutions, results influenced by parameter and initialization

3.2 Comparison of dimension reduction results on batch data

This section explains how NMF, MCR and PCA results quality can be evaluated on spectral data organized in multiple batches. When the data structure consists of batches, it is possible to train models individually on each batch, globally on all batches, or on part of the batches and apply the trained model on the batches left out. Considering an industrial context, the third approach, called in this article the "leave-one-batch-out model", would help for instance to visualize the evolution of a new batch based on a model from historical batches. This section presents these 3 approaches and in section 4.1, these concepts will be applied on pharmaceutical data and compared in terms of interpretability and model fit.

3.2.1 Individual model

The first approach consists in applying the decomposition to each matrix $\mathbf{X}_i$ corresponding to the spectra of a single batch or experiment i of size $t_i \times m$.

Each matrix can then be decomposed into $\mathbf{X}_i = \mathbf{H}_i \mathbf{W}'_i + \mathbf{E}_i$ and the error of the model is calculated as $\mathbf{E}_i = \mathbf{X}_i - \mathbf{H}_i \mathbf{W}'_i$.

The quality of fit can then be quantified by the percentage of error as $\%error_i = \frac{\|\mathbf{E}_i\|_F^2}{\|\mathbf{X}_i\|_F^2} \times 100$ and the mean percentage of error by $\frac{1}{I} \sum_{i=1}^I \%error_i$

3.2.2 Global model

The second approach consists in applying the decomposition on the global matrix $\mathbf{X}_G$ of size $T \times m$ where $T = \sum_{i=1}^I t_i$. This matrix corresponds to the spectra of all experimental matrices $\mathbf{X}_i$, with $i = 1 \rightarrow I$, stacked vertically.

The global matrix can then be decomposed into $\mathbf{X}_G = \mathbf{H}_G \mathbf{W}'_G + \mathbf{E}_G$ and the error of the model for each submatrix i calculated as $\mathbf{E}_{G_i} = \mathbf{X}_i - \mathbf{H}_{G_i} \mathbf{W}'_G$.

The percentage of error for one experiment can then be calculated as $\%error_i = \frac{\|\mathbf{E}_{G_i}\|_F^2}{\|\mathbf{X}_i\|_F^2} \times 100$.

3.2.3 Leave-one-batch-out model

The third approach consists in applying the decomposition to the matrices $\mathbf{X}_{G/i}$ of size $(T - t_i) \times m$ corresponding to the spectra of all batches $\mathbf{X}_G$ without one experiment $\mathbf{X}_i$. Each matrix can be decomposed into $\mathbf{X}_{G/i} = \mathbf{H}_{G/i} \times \mathbf{W}'_{G/i} + \mathbf{E}_{G/i}$.

Using obtained loadings $\mathbf{W}_{G/i}$ and spectra $\mathbf{X}_i$ of the batch left out, scores $\hat{\mathbf{H}}_i$ are calculated by projecting spectra $\mathbf{X}_i$ on the space of $\mathbf{X}_{G/i}$ components. The calculation is done using least squares: $\hat{\mathbf{H}}_i = (\mathbf{X}_i \mathbf{W}_{G/i}) \times (\mathbf{W}'_{G/i} \mathbf{W}_{G/i})^{-1}$.

The percentage of error for each submatrix $\mathbf{X}_i$ is then calculated as $\%error_i = \frac{\|\mathbf{X}_i - \hat{\mathbf{H}}_i \mathbf{W}'_{G/i}\|_F^2}{\|\mathbf{X}_i\|_F^2} \times 100$.

3.2.4 The special case of PCA

In PCA, above formulae need to be a little detailed since a centering (and possibly a scaling) is, in most cases, applied before dimension reduction. It is indeed important to emphasize which data should be used for centering in the three approaches.

With the individual model, each $\mathbf{X}_i$ will be column-centered by the matrix $\mathbf{M}_i$ containing the means of each column of $\mathbf{X}_i$ such that $\mathbf{X}_i^c = \mathbf{X}_i - \mathbf{M}_i$. PCA decomposition becomes then $\mathbf{X}_i^c = \mathbf{H}_i^c \mathbf{W}_i^{c'} + \mathbf{E}_i^c$. The percentage of error for each experiment is then defined as $\%error_i = \frac{\|\mathbf{E}_i^c\|_F^2}{\|\mathbf{X}_i^c\|_F^2} \times 100$.

With the global model, $\mathbf{X}_G$ is centered with $\mathbf{M}_G$ containing the means of each column of $\mathbf{X}_G$: $\mathbf{X}_G^c = \mathbf{X}_G - \mathbf{M}_G$. PCA decomposition becomes $\mathbf{X}_G^c = \mathbf{H}_G^c \mathbf{W}_G^{c'} + \mathbf{E}_G^c$ and the percentage of error for each submatrix is $\%error_i = \frac{\|\mathbf{E}_i^c\|_F^2}{\|\mathbf{X}_i^c\|_F^2} \times 100$.

With the leave-one-batch-out model, the training set $\mathbf{X}_{G/i}$ is first centered by $\mathbf{M}_{G/i}$: $\mathbf{X}_{G/i}^c = \mathbf{X}_{G/i} - \mathbf{M}_{G/i}$. The prediction set $\mathbf{X}_i$ is then column-centered by the matrix $\hat{\mathbf{M}}_i$ of size $(t_i \times m)$ containing the means of the training set $\mathbf{M}_{G/i}$ such that $\mathbf{X}_i^c = \mathbf{X}_i - \hat{\mathbf{M}}_i$. Scores $\hat{\mathbf{H}}_i^c$ are then calculated simply as $\hat{\mathbf{H}}_i^c = \mathbf{X}_i^c \times \mathbf{W}_{G/i}^c$ and the percentage of error $\%error_i = \frac{\|\mathbf{X}_i^c - \hat{\mathbf{H}}_i^c \mathbf{W}_{G/i}^{c'}\|_F^2}{\|\mathbf{X}_i^c\|_F^2} \times 100$.

Note that for each of those three cases, eventough the centering is advantageous for PCA, the error percentages $\%error_i$ are compared against $\mathbf{X}_i$ not centered in order to allow comparion with NMF and MCR.

3.3 Multivariate calibration method PLSR

PLSR is a multiple linear regression technique aimed at predicting/explaining a vector $\mathbf{y}$ of dimensions $T \times 1$ (PLS1) or a matrix $\mathbf{Y}$ of dimensions $T \times n$ (PLS2) from the columns $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$ of a matrix $\mathbf{X}$. Generic notations $\mathbf{Y}$ and $\mathbf{X}$ will be used but those matrices could correspond to global matrices $\mathbf{Y}_G^*$ and $\mathbf{X}_G$, as well as individual matrices $\mathbf{Y}_i^*$ and $\mathbf{X}_i$ described in the section 2.1. Contrarily to classical ordinary least squares (OLS) regression, PLS can handle data matrices $\mathbf{X}$ with a number of columns m (variables) higher than the number of rows T (observations/samples). This subsection is devoted to PLS2 and uses the following notations:

X	$T \times m$	data matrix of explanatory variables (column centered and possibly scaled)
Y	$T \times n$	response matrix to be explained by X . Y is column centered and possibly scaled
h_r	$T \times 1$	X scores vector r ($r : 1, \dots, R$) with $R \leq \min(T, m)$
H	$T \times R$	matrix of X scores vectors
w_r	$m \times 1$	X loadings vector r ($r : 1, \dots, R$)
W	$m \times R$	matrix of X loading vectors
u_r	$T \times 1$	Y scores vector r ($r : 1, \dots, R$)
U	$T \times R$	matrix of Y scores vectors
q_r	$n \times 1$	Y loadings vector r ($r : 1, \dots, R$)
Q	$n \times R$	matrix of Y loading vectors
E_x	$T \times m$	X decomposition residual matrix
E_y	$T \times n$	Y decomposition residual matrix
E	$T \times n$	matrix of prediction model residuals
B	$m \times n$	matrix of model coefficients

The primary goal of PLS regression is to find a linear relation $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ between the **x**- and **y**- variables using an $m \times n$ coefficient matrix **B** and a prediction error matrix **E**.

This relation is not found directly since matrices **X** and **Y** are first decomposed into score matrices **H** and **U**, loading matrices **W** and **Q** and residuals **E_x** and **E_y**:

$$\mathbf{X} = \mathbf{HW}' + \mathbf{E}_x$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{E}_y$$

In PLS, the covariance between the corresponding **X** and **Y** score vectors is maximized:

$$\max \text{cov}(\mathbf{h}_r, \mathbf{u}_r)^2 = \max \text{corr}(\mathbf{h}_r, \mathbf{u}_r)^2 \text{var}(\mathbf{h}_r) \text{var}(\mathbf{u}_r)$$

Maximizing the covariance is equivalent to maximize the correlation between the **X** and **Y** scores AND maximize the variance of the **X** and **Y** scores which ensures the quality of the dimension reduction in both **X** and **Y** spaces.

The orthogonal properties of the scores and loadings depend on the PLS algorithm. NIPALS algorithm was the first algorithm for solving the PLS problem but other very known algorithms are for example SIMPLS and Kernel algorithms (Rosipal & Krämer, 2005; Varmuza & Filzmoser,

2016).

The way to obtain the matrix of model coefficients $\mathbf{B}$, which will be noted $\mathbf{B}^{PLS}$ in the rest of this article, depends on the algorithm which is used but the final coefficients are identical. It can also be noted that $\mathbf{X}$ and $\mathbf{Y}$ are in most cases centered and sometimes standardized before the application of PLS.

3.4 Prediction model set-up and comparison

This section presents two methodologies to build prediction models of the concentrations of chemical compounds from the spectral and UPLC data. The "one-step leave-one-batch-out PLS" and the "two-step leave-one-batch-out regression". Goodness of fit measure are then provided to allow their comparison.

Considering an industrial context, such methodologies can be used to build a model from a set of historical batches, and then apply the model on a new batch to get predictions in real-time without sampling the off-line measurements. The one-step methodology uses PLS to fit directly the UPLC data to spectral data. With the two-step methodology, a dimension reduction is first applied on the spectra, e.g. by PCA, NMF or MCR, and then a multivariate regression model is then built to predict UPLC data from the obtained scores and UPLC data. This approach is aimed to enhance the interpretability and robustness of the models.

3.4.1 One-step PLS approach

Let's consider the off-line matrix $\mathbf{Y}_G^*$ of dimensions $S \times n$ where $S = \sum_{i=1}^I s_i$, s_i being the number of rows of each subset matrix $\mathbf{Y}_i^*$ corresponding to experiment i . Let's also consider the matrix $\mathbf{X}_G^*$ defined in section 2.1 which is a subset of $\mathbf{X}_G$ with the spectra corresponding to off-line sampling times. $\mathbf{X}_G^*$ is of dimensions $S \times m$, i.e. with the same number of rows as $\mathbf{Y}_G^*$.

The idea of the one-step leave-one-batch-out PLS is to build a PLS model based on $\mathbf{X}_{G/i}^*$ $((S - s_i) \times m)$ and $\mathbf{Y}_{G/i}^*$ $((S - s_i) \times n)$. $\mathbf{X}_{G/i}^*$ and $\mathbf{Y}_{G/i}^*$ being respectively the reduced spectral matrix of all batches, without the i^{th} batch $\mathbf{X}_i^*$ $(s_i \times m)$, and the off-line matrix of all batches without $\mathbf{Y}_i^*$ $(s_i \times n)$. The resulting PLS prediction model is then written as:

$$\hat{\mathbf{Y}}_{G/i}^* = \mathbf{X}_{G/i}^* \hat{\mathbf{B}}_{G/i}^{PLS}$$

3.4.2 Two-step dimension reduction + regression approach

1. In the first step, a dimension reduction is applied on matrix $\mathbf{X}_{G/i}$ (dimensions $(T - t_i) \times m$) of all spectra without experiment i , to obtain the scores $\mathbf{H}_{G/i}$ and the loadings $\mathbf{W}_{G/i}$ by PCA, NMF or MCR:

$$\mathbf{X}_{G/i} = \mathbf{H}_{G/i} \times \mathbf{W}_{G/i}' + \mathbf{E}_{G/i}$$

$\mathbf{H}_{G/i}^*$ is then defined as the submatrix of $\mathbf{H}_{G/i}$ with the scores corresponding to the offline data.

2. In the second step, a multivariate regression model is defined to link the spectral data scores $\mathbf{H}_{G/i}^*$ to UPLC data $\mathbf{Y}_{G/i}^*$:

$$\mathbf{Y}_{G/i}^* = \mathbf{H}_{G/i}^* \mathbf{C}_{G/i} + \mathbf{E}$$

Coefficients $\mathbf{C}_{G/i}$ and scores $\mathbf{H}_{G/i}^*$ are estimated by least squares such as $\hat{\mathbf{C}}_{G/i} = (\mathbf{H}_{G/i}^{*'} \mathbf{H}_{G/i}^*)^{-1} \times (\mathbf{H}_{G/i}^{*'} \mathbf{Y}_{G/i}^*)$. Now, since $\hat{\mathbf{H}}_{G/i}^* = \mathbf{X}_{G/i}^* \mathbf{W}_{G/i} \times (\mathbf{W}_{G/i}' \mathbf{W}_{G/i})^{-1}$ (see section 3.2.3), the resulting model equation can also be written as:

$$\mathbf{Y}_{G/i}^* = \mathbf{H}_{G/i}^* \mathbf{C}_{G/i} + \mathbf{E} = \mathbf{X}_{G/i}^* \mathbf{W}_{G/i} \times (\mathbf{W}_{G/i}' \mathbf{W}_{G/i})^{-1} \mathbf{C}_{G/i} + \mathbf{E}$$

A two-step regression coefficient matrix $\hat{\mathbf{B}}_{G/i}^{2S}$ can then be defined as $\hat{\mathbf{B}}_{G/i}^{2S} = \mathbf{W}_{G/i} \times (\mathbf{W}_{G/i}' \mathbf{W}_{G/i})^{-1} \hat{\mathbf{C}}_{G/i}$ and the resulting two-step prediction model summarized as:

$$\hat{\mathbf{Y}}_{G/i}^* = \mathbf{X}_{G/i}^* \hat{\mathbf{B}}_{G/i}^{2S}$$

3.4.3 Using the model

The model estimated on the training set can then be used to predict the full profiles of the chemical concentrations over time and visualize off-line values directly from the spectra.

For the training set, all intermediate off-line values, which were not sampled, can be predicted and represented graphically from the full spectral matrix $\mathbf{X}_{G/i}$ as $\hat{\mathbf{Y}}_{G/i} = \mathbf{X}_{G/i} \hat{\mathbf{B}}_{G/i}$, with $\hat{\mathbf{B}}_{G/i}$ being $\hat{\mathbf{B}}_{G/i}^{PLS}$ or $\hat{\mathbf{B}}_{G/i}^{2S}$.

Then, off-line values for the batch left out $\mathbf{X}_i$ (or any new batch) can be predicted as $\hat{\mathbf{Y}}_i = \mathbf{X}_i \hat{\mathbf{B}}_{G/i}$.

Finally, predictions of the measured off-line values can be extracted from $\hat{\mathbf{Y}}_i$ as $\hat{\mathbf{Y}}_i^*$ in order to evaluate the quality of the model. A cross validation root mean square error for experiment i can be defined as:

$$RMSECV_{ij} = \sqrt{\frac{\sum_{k=1}^{s_i} (\mathbf{Y}_{ijk}^* - \hat{\mathbf{Y}}_{ijk}^*)^2}{s_i}}$$

with $j : 1, \dots, n$ corresponding to the response columns of $\mathbf{Y}_G^*$, and $k : 1, \dots, s_i$ to the number of rows of the subset matrix $\mathbf{Y}_i^*$ for batch i .

A global RMSECV can be calculated as $RMSECV_j = \frac{1}{I} \sqrt{\frac{\sum_{k=1}^{s_i} (\mathbf{Y}_{ijk}^* - \hat{\mathbf{Y}}_{ijk}^*)^2}{s_i}}$.

It can also be noted that the reasoning of section 3.2.4 about centering can be applied to one-step PLS and two-step data reduction + regression approach.

4 Case study

This section uses the dataset described in section 2.1 to illustrate how the techniques presented in section 3 can be used for 2 purposes. First to visualize and interpret chemical reactions, and second to predict the concentrations of chemical components.

4.1 Visualization and interpretation of chemical reactions

4.1.1 R packages implementation

R software was used for the implementation of the data reduction techniques PCA, NMF and MCR. This section provides details on this implementation.

For PCA, the function `SVDforPCA` from the R package *MBXUCL* (Lapointe-Garant et al., 2008) was used. Centering was applied before singular value decomposition but no standardization.

For NMF, the function `spectralNMF` from the R package *spectralAnalysis* (version 3.12.0, (Van Oirbeek et al., 2018)) was used. Random initialization as well as initialization using prior loadings have been tested. For random initialization, $\mathbf{H}_0$ and $\mathbf{W}_0$ were set to pseudo-random nonnegative values between 0 and 1. The NMF algorithm PGNMF was used (Lin, 2007). The stopping criterion was set such that if $(\|\mathbf{E}_{k+1}\|_F - \|\mathbf{E}_k\|_F) \leq 0.00001$ the algorithm stopped.

For MCR, the function `als` from the R package *ALS* (version 0.0.6, (Mullen, 2015)) was used. Random initialization as well as initialization using prior loadings were also tested with the same approach as for NMF. The algorithm used for the `als` function is alternating least square. The stopping criterion was set such that if $\frac{\|\mathbf{E}_{k+1}\|_F - \|\mathbf{E}_k\|_F}{\|\mathbf{E}_k\|_F} \leq 0.001$ the algorithm stopped.

4.1.2 Results

This section shows the results of the application of dimension reduction methods PCA, MCR and NMF to the matrix $\mathbf{X}_G$ and submatrices $\mathbf{X}_i$ following the individual, global or leave-one-batch-out methodologies. Error percentages for each batch and methodology were computed in order to compare the fit of the different models, and the scores and loadings were analyzed to compare the interpretability of each method.

Figure 2 shows the evolution of the mean error percentage when the number of components in the model increases. Errors percentages are presented for PCA, MCR and NMF following the

individual, global or leave-one-batch-out methodologies up to 6 components. For NMF and MCR the convergence of the algorithm started only at 2 components. It can be observed that the mean error percentage decreases when the number of components increases and that errors tend to be closer for PCA, NMF and MCR with more components. For the global and leave-one-batch-out methodologies, the slopes of the decrease start to be similar at 3 components, whereas for the individual methodology 4 components seem to be necessary. For the individual model, a shift to the left of PCA compared to NMF and MCR can be observed for the first few components. This suggest that PCA requires one component less to reach the same error percentage as NFM and MCR. One reason could be that PCA needs only 1 component to explain the reactant and product correlated negatively, while NMF and MCR requires 2 components. It can also be observed that the mean error decreases differently for the individual methodology, whereas it is similar for the global and leave-one-batch-out methodologies. The main goal of this figure is to guide the decision for the number of components to use in the model. In this case, the choice was difficult between 2 and 3 components, and there was a trade-off to be chosen between the fit and interpretability. In concertation with chemists, and to improve interpretability, two components were finally chosen for the calculation of model errors as well as to display scores and loadings. Extra components were indeed not correctly modeled by NMF and MCR and reduced the interpretability of the first two components modelizing the reactant and the product. Results with 3 components are nevertheless given in appendix.

Error percentages for each batch are given by methodology in Table III. Moreover, percentages for NMF/MCR with random priors are also given.

Looking to the mean percentages of error per column, it can first be observed that, for every method (PCA, NMF or MCR), lowest errors are obtained using the individual methodology, followed by the global and leave-one-batch-out approaches. This can be explained since the individual methodology finds loadings and scores for each batch separately, leading to results fitted for each batch. The global methodology finds loadings common to all batches, and next, the scores are extracted per experiment. Those loadings have the advantage to be common for all experiments but lead to less fitted results than the individual methodology. And finally, the leave-one-batch-out methodology finds common loadings to all batches, except the batch left out. The scores are then estimated for the batch left out based on the loadings found for other experiments, leading to higher error percentages than individual and global methodologies.

It can then be seen that model errors of NMF and MCR, with and without priors, are similar for each methodology (individual, global and leave-one-batch-out). Those results are quite expected

Figure 2: Mean error percentages by number of components.

since even if the algorithms used for MCR and NMF are different, the optimized criterion is the same and results are logically similar. With random priors, iterations were also giving similar results both for NMF and MCR.

Looking at the errors for individual models per batch for PCA, experiment 5 has the highest error percentage, followed by experiments 4, 8, 2 and 11. For NMF and MCR, the highest error percentages are obtained for experiment 8, 11, 5 and 2.

Looking at errors for global models, it can be seen that for PCA, it is experiment 2 which has the highest error, which is also the case for methods NMF and MCR.

	PCA centered			MCR			MCR random priors			NMF			NMF random priors		
	Individual	Global	LoBo	Individual	Global	LoBo	Individual	Global	LoBo	Individual	Global	LoBo	Individual	Global	LoBo
<i>Batch number</i>															
1	0.006	0.292	0.348	0.013	0.379	0.436	0.013	0.379	0.435	0.013	0.375	0.436	0.013	0.373	0.433
2	0.016	1.182	1.821	0.101	1.942	2.413	0.101	1.941	2.413	0.098	1.950	2.422	0.098	1.937	2.417
3	0.007	0.228	0.384	0.016	0.192	0.202	0.016	0.193	0.202	0.016	0.176	0.205	0.016	0.173	0.201
4	0.019	0.424	0.501	0.062	0.364	0.416	0.062	0.364	0.416	0.063	0.363	0.417	0.062	0.361	0.414
5	0.048	0.284	0.709	0.125	0.411	0.881	0.125	0.411	0.880	0.124	0.414	0.888	0.124	0.411	0.882
6	0.004	0.095	0.121	0.022	0.197	0.223	0.022	0.198	0.223	0.021	0.188	0.222	0.021	0.188	0.222
7	0.001	0.143	0.163	0.010	0.165	0.181	0.010	0.165	0.181	0.010	0.162	0.183	0.010	0.160	0.181
8	0.018	0.355	0.534	0.144	0.552	0.707	0.144	0.551	0.707	0.141	0.544	0.701	0.141	0.550	0.710
9	0.003	0.119	0.356	0.023	0.182	0.246	0.023	0.183	0.246	0.023	0.162	0.246	0.023	0.162	0.245
10	0.003	0.060	0.068	0.012	0.073	0.078	0.012	0.073	0.078	0.012	0.072	0.081	0.012	0.069	0.078
11	0.008	0.636	0.718	0.143	0.749	0.844	0.143	0.749	0.844	0.142	0.744	0.846	0.142	0.742	0.843
Mean	0.012	0.347	0.520	0.061	0.473	0.602	0.061	0.473	0.602	0.060	0.468	0.604	0.060	0.466	0.603

Table III: Model error percentages per batch and methodology.

Figure 3 presents the loadings obtained by PCA, NMF and MCR from the individual models applied on experiment 1. Sub-figure A shows the loadings of PCA for the first two principal components, sub-figure B shows the loadings obtained by NMF and MCR corresponding to the reactant and product, and sub-figure C shows the pure spectra of the reactant and product obtained by MIR spectroscopy. The pure spectra were used to initialize NMF and MCR algorithms which allows to make the correspondence among components and chemical compounds. Note that random initialization was also used and provided similar results. It can first be observed that PCA loadings have positive and negative values which makes the correspondence with spectral profiles not straightforward. Indeed, components calculated by PCA are less designed to capture the chemical compounds profiles. In the case of NMF and MCR, the loadings have only positive values, due to the non-negativity constraint, and a correspondence among components and chemical compounds is expected. One can see in Figure 3 that the loadings are not exactly the same as the spectral profiles but are easier to interpret for the chemists than in the case of PCA. It can finally be noted that the loadings found by NMF and MCR are extremely close, meaning that those two methods did converge to similar solutions.

Figure 3: Loadings for experiment 1 based on individual models.

Figure 4 presents the scores obtained by PCA, NMF and MCR from the individual model applied on experiment 1. Sub-figure A shows the scores of PCA for the first two principal components, sub-figure B shows the scores obtained by NMF and MCR corresponding to the reactant and product, and sub-figure C shows the profiles of the reactant and product of the reaction obtained by UPLC. In sub-figure C, the dots correspond to measured values and the dashed line simply relies them. Those UPLC profiles can be used as reference. It can first be observed that PCA scores have positive and negative values which makes the interpretation not straightforward. The first component might correspond to both the evolution of the product and the reactant since they are negatively correlated, but the second one does not have a clear pattern. In the case of NMF and MCR, the scores have only positive values and correspond to the evolution of the quantities of the chemical compounds of the reaction, which is easier for interpretation. The quantity of the reactant decreases over time and the product increases such as seen with UPLC values. The scale of NMF and MCR scores differ of course from UPLC values but, for monitoring purpose, only the observation of the kinetics trends is important. Finally, it can be noted that scores given by NMF and MCR are also quite close but not identical.

Figure 4: Scores for experiment 1 based on individual models.

Figure 5 presents the scores obtained by PCA for individual models. Each plot shows the scores of the first against the second principal component for each experiment. Moreover, the line connecting the values has been colored from green to red, to indicate at which time of the reaction the scores correspond. Percentages of variation are also indicated for the first two components. For the first principal component, the percentages range between 84.5 and 99.2%. For the second principal component, the percentages range between 0.5 and 9.5%. It can be seen that experiment 4 presents many fluctuations. Those fluctuations were also detected by NMF on the score plots of the starting material and product (Fig. 6). Experiment 4 is however not the one with the highest error percentage (0.019%), but experiment 5 as observed previously in Table III (0.048%). It can be seen that experiments 2, 5, 8 and 11 have all an irregular pattern detected on the second component at the end of the reaction. Those 4 experiments have also higher error percentages than others, with respectively 0.016, 0.048, 0.018 and 0.008 %. Interestingly, the fluctuations seen on experiments 2, 5, 8 and 11 were not observed on the score plots of NMF (Fig. 6). This might suggest that those fluctuations would not be related to chemical components but rather to another factor captured by the second principal component of PCA.

Figure 5: Scores 1 vs scores 2 values for each experiment based on individual models.

Figure 6 presents the scores obtained by NMF for the reactant and product based on the global model (sub-figure A) as well as the UPLC values (sub-figure B). The global model allows to visualize all the reactions on a same space and to compare their kinetics. It was chosen to display only NMF scores since MCR provides similar results. Similarly to the individual models, the pure spectra of the reactant and product were used to initialize the NMF algorithm. It can be observed that all scores have positive values making the interpretation easier. The reactant scores decrease and the product scores increase for each experiment as expected. Important differences can be observed in the conversion speed among the reactions. Indeed, those reactions come from a research and development context and were performed on extreme conditions to understand the impact of different process parameters. Experiment 5, followed by 2 and 8 present the fastest kinetics. The product increases really fast during the first hour, is then produced linearly, until reaching a plateau. The plateau is reached around 5 hours for experiment 5, and around 10 hours for experiments 2 and 8. It is also interesting to see an irregular pattern on the second principal component of those experiments on the PCA score plots (see Fig. 5). On the opposite, experiments 3 and 9 were among the slowest, and did not reach a plateau even after 21 hours. Finally, experiment 4 presents a few fluctuations as seen before with PCA. The advantage of NMF compared to UPLC is clearly seen on this figure. Indeed, since spectral measurements are taken continuously, NMF allows to visualize the complete kinetics trend, and to detect fluctuations or plateaus in the reaction. For UPLC, since only four samples were taken during the reaction, some information can be missed.

Figure 6: UPLC values and NMF scores over time based on the global model.

A few practical points and results can be emphasized from this case study:

- Chemical reactions can be visualized with individual and global models. Since the dimension-reduction of individual models is done on different spaces, those models are more appropriate to visualize experiments separately than to compare the trends together. They can be used for example in a first step as a benchmark when there are only a few experiments or when the chemical reactions are different. In the case of global models, the dimension-reduction is done on a same space for all experiments, which allows to visualize and compare them together. Global models present moreover the advantage of being more robust, since several experiments are used to build the model, which can prevent over-fitting. The space of a global model can also be used to project in real time the spectra of new batches and visualize directly how they perform compared to the experiments used to build the model. This can be for example useful to control if a new experiment stays in a certain range or differs from it.
- NMF and MCR provide an interpretable visualization of chemical reactions. Indeed, those methods allow to choose the number of components upfront to correspond to the number of chemical compounds. Moreover, the loadings are closer to the pure spectra and the scores closer to UPLC profiles due to the non-negativity constraint. Eventhough PCA is

less interpretable, applying it remains a good tool in this context since it can help to detect particular behaviours in reactions (e.g. experiment 4) and take the decision to stop them if necessary. Other examples are experiments 2, 5 and 8 which presented an irregular pattern on the second PCA component, at the end of the reaction, and those experiments were seen as having the fastest kinetics in NMF.

- NMF and MCR provide equivalent results eventough their algorithms are different. The only constraint applied was non-negativity since the absorbance of the spectra and the concentration of the chemical compounds are supposed to be negative. However, it is important to keep in mind that other constraints can be applied depending of the chosen package. This is well known for MCR, which was developed to solve chemical mixture problems, and for which the algorithms propose constraints such as closure and unimodality.

4.2 Chemical components prediction models

Section 4.1 was focusing on how to visualize and interpret the kinetics of chemical reactions using methods PCA, NMF and MCR on spectral data. Those methods can be useful to follow the trend of a chemical reaction over time but do not allow to quantify the concentration of the chemical components.

This section illustrates the use of PLS and multivariate regression to predict the concentrations of chemical components using the one- and two-step leave-one-batch-out methodologies presented in section 3.4.

4.2.1 Results

Figure 7 shows the evolution of the mean RMSECV of the product for one-step PLS, and two-step multivariate regression based on the PCA, NMF and MCR scores, when the number of components in the model increases. For PLS, the mean RMSECV decreases from 1 to 2 components and then stays stable. For PCA, NMF and MCR, the mean RMSECV's globally decrease from 2 to 5 components to get closer to the value of PLS. As for Figure 2, the idea of this figure is to guide the decision for the number of components to use in the model. In this case, the choice was again to use 2 components. Indeed, for NMF and MCR only the first two components modelize correctly the chemical compounds and there was not a big gain to use 3 components.

Table IV presents the RMSECV of the product for each batch obtained by the one-step PLS with 2 components, and two-step multivariate regression based on the scores of PCA, NMF and MCR also with 2 components. For the two-step multivariate regression, only two components were used since only those are meaningful in NMF and MCR. For comparability, the results with the first two components for PCA and PLS were also used.

It can be observed that in average, by methodology, the lowest RMSECV is obtained for the one-step PLS (5.12), two-step regression + PCA (5.71), two-step regression + NMF (6.03) and finally the two-step regression + MCR (7.06). Those results show that the one-step PLS provides slightly lower RMSECV's than the two-step methodology. However the scores obtained by PCA, MCR and NMF can also be used for prediction, with the advantage for MCR and NMF to be interpretable. Those RMSECV values should be put into perspective. Indeed, the data used in this article were not designed specifically to build a calibration model, so it would be possible to reach lower RMSECV's with a better design. As a point of comparison it can also be mentioned that UPLC values range from 0 to 100, which shows that the differences between the one- and two-step models are relatively small. It can also be pointed out that the one-step PLS is not always the best. In-

Figure 7: Mean RMSECV of the product by number of components.

Table IV: RMSECV of the product from one-step PLS and two-step multivariate regression.

Batch	One-step PLS	Two-step PCA	Two-step MCR	Two-step NMF
1	1.95	2.84	4.62	3.34
2	4.64	5.12	4.62	5.66
3	7.02	7.26	9.45	9.03
4	10.10	10.39	8.44	10.07
5	12.17	12.36	13.15	12.47
6	5.53	6.53	10.05	6.77
7	0.81	1.46	3.36	1.76
8	4.65	4.52	2.79	4.47
9	4.44	6.39	14.46	6.94
10	1.48	2.15	3.53	2.24
11	3.48	3.83	3.22	3.62
Mean	5.12	5.71	7.06	6.03

deed, for some batches, the two-step approach provides lower RMSECV's. This is for example the case for batch 2 (MCR), batch 4 (MCR, NMF), batch 8 (PCA, MCR, NMF), and batch 11 (MCR).

Figure 8 presents the predictions for experiment 1 based on the one-step leave-one-batch-out PLS with 2 components, and the two-step leave-one-batch-out regression, based on PCA, MCR and NMF with also two components. This means that the models were built on all experiments except

experiment 1, and then applied on the spectra and scores of this experiment. The lines represent the predicted values whereas the points represent the observed UPLC values. One can see that the best fit is obtained for the one-step PLS, whereas the least good is for the two-step regression based on MCR scores as seen on Table IV. It is also important to note that the predictions are obtained at all intermediate points not sampled by UPLC, which allows to visualize continuously the reaction. When looking at the kinetics, it can be observed that the reaction is the fastest during the first hour, then continues constantly up to 21 hours, but does not reach a plateau.

Figure 8: Leave one batch out PLS prediction for batch 1.

From a practical point of view, this section shows that spectra as well as scores can be used to build a calibration model when reference measurements are available. Three applications are possible. The first one is to build a PLS model on a single experiment to visualize it afterwards. As said before, it can allow to interpolate chemical concentrations at all time points which were not sampled off-line. This can help to visualize the complete kinetics or particular patterns. The second application is to build a PLS model on multiple experiments. The visualization is also done afterwards but this kind of model has the advantage of being more robust than for one experiment and allows the comparison of multiple experiments together. And finally, as shown with Figure 8, the third application is to use a global model to predict the chemical concentrations over all

time points of a new reaction instead of taking off-line measurements. Those results show that with 10 experiments and 4 sampling times per experiment, predicting a reaction of 21 hours is possible. The kinetics of this reaction was of course simple, and only two compounds needed to be visualized, but those are coming from a realistic industrial context, and the principle stays the same for more complex reactions. Two recommendations can be given to build this kind of model in a research and development context. When the goal is to understand the impact of different factors on the reaction, the first recommendation is to generate data from a design of experiment (DoE). Indeed DoE's can help to capture the effect of the different factors with an optimal number of experiments. Since those reactions occur over time, the second recommendation is to sample off-line measurements at enough time points in order to correctly capture the kinetics of the reaction.

5 Conclusion and perspectives

Following the Process Analytical Technology initiative promoted by the Food and Drug Administration, spectroscopic methods have become popular in the pharmaceutical industry to continuously monitor critical quality attributes during chemical processes. Spectroscopic methods generate however multivariate data requiring the use of chemometrics approaches to be interpretable.

This paper used a set of chemical experiments performed in an industrial research and development context to study the synthesis of an active molecule. Those experiments were monitored on-line by mid-infrared spectroscopy and off-line by UPLC.

This paper presented a practical framework to answer two objectives. The first objective was to compare chemometrics approaches PCA, NMF and MCR to visualize, monitor and interpret chemical reactions. The second objective was to propose different approaches to build real-time prediction models to predict chemical concentrations using calibration methods PLS and multivariate regression.

For the first objective, it has been seen that methods NMF and MCR allowed a better interpretation of spectral data than PCA thanks to the non-negativity constraint and the prior choice of the number of components. Results also showed that even though NMF and MCR come from different fields, they share the same objectives and provide in practice similar results.

For the second objective, it has been seen that spectra or scores obtained by PCA, NMF and MCR could be used to build a calibration model with UPLC measurements. A one-step approach using the spectra and a two-step approach using the scores were compared and provided close results. In practice, those calibration models allow to predict in real time, and continuously, the chemical concentrations of a new reaction without having to sample off-line measurements. Broader than the statistical comparisons, this paper intended to present different tools to assist scientists in the development of chemical compounds.

Acknowledgements

The authors thank Janssen Pharmaceutica and the Université catholique de Louvain for their support. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is also gratefully acknowledged.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Softwares

The R software environment was exclusively used (<http://www.R-project.org>).

Conflict of Interest

Authors declare that they have no conflict of interest.

Compliance with ethical requirements

This study analyses collected data which are not involving human participants.

References

- Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1), 155–173.
- Boutsidis, C., & Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4), 1350–1362. <https://doi.org/https://doi.org/10.1016/j.patcog.2007.09.010>
- Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons.
- De Beer, T., Burggraef, A., Fonteyne, M., Saelens, L., Remon, J. P., & Vervaet, C. (2011). Near infrared and raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *International journal of pharmaceutics*, 417(1-2), 32–47.

- Erichson, N. B., Voronin, S., Brunton, S. L., & Kutz, J. N. (2016). Randomized matrix decompositions using r. *arXiv preprint arXiv:1608.02148*.
- Gillis, N. (2014). The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 257–291.
- ICH Guidance for Industry. (2009). Ich harmonised tripartite guideline q8(r2) - pharmaceutical development.
- Langville, A. N., Meyer, C. D., Albright, R., Cox, J., & Duling, D. (2014). Algorithms, initializations, and convergence for the nonnegative matrix factorization. *arXiv preprint arXiv:1407.7299*.
- Lapointe-Garant, P.-P., Simard, J.-S., & Abatzoglou, N. (2008). Real-time nir monitoring of a pharmaceutical blending process through multivariate analysis-derived models. (7).
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788.
- Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10), 2756–2779.
- Loureço, N., Lopes, J., Almeida, C., Sarraguça, M., & Pinheiro, H. (2012). Bioreactor monitoring with spectroscopy and chemometrics: A review. *Analytical and bioanalytical chemistry*, 404(4), 1211–1237.
- Mullen, K. M. (2015). *Als: Multivariate curve resolution alternating least squares (mcr-als)*. Retrieved November 27, 2020, from <https://cran.r-project.org/web/packages/ALS/index.html>
- Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111–126. <https://doi.org/10.1002/env.3170050203>
- Rosipal, R., & Krämer, N. (2005). Overview and recent advances in partial least squares, 34–51.
- Ruckebusch, C. (2016). *Resolving spectral mixtures: With applications from ultrafast time-resolved spectroscopy to super-resolution imaging* (Vol. 30). Elsevier.
- Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization, 540–547.
- Tauler, R., & Maeder, M. (2009). 2.20 - two-way data analysis: Multivariate curve resolution – error in curve resolution. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics* (pp. 345–363). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-044452701-1.00051-X>

- Tauler, R., Smilde, A., & Kowalski, B. (1995). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics*, 9(1), 31–58. <https://doi.org/10.1002/cem.1180090105>
- Van Oirbeek, R., Blommaert, A., Sauwen, N., Maes, T., Dijkmans, J., Cuypers, J., Khamiakova, T., Thiel, M., & Beleites, C. (2018). *Spectralanalysis: Pre-process, visualize and analyse process analytical data, by spectral data measurements made during a chemical process*. Retrieved December 20, 2020, from <https://cran.r-project.org/web/packages/spectralAnalysis/index.html>
- Varmuza, K., & Filzmoser, P. (2016). *Introduction to multivariate statistical analysis in chemometrics*. CRC press.

Appendix

Figure 9 presents the scores obtained by NMF for the reactant, product and a side product based on the global model (sub-figure A) as well as the UPLC values (sub-figure B). It can first be observed that the side product is not well modelled and is too high compared to the main product. A second observation is that the main product is not modelled as well as in Figure 6. For exemple, experiments 2 and 5 reach lower values at the end of the reaction than what is observed with UPLC values.

Table V shows the error percentages for each batch and methodology with 3 components. Those errors can be compared with table III.

Figure 9: UPLC values and NMF scores over time based on the global model.

	PCA centered			MCR			MCR random priors			NMF			NMF random priors		
	Individual		Global	Individual		Global	Individual		Global	Individual		Global	Individual		Global
	LoBo	LoBo		LoBo	LoBo		LoBo	LoBo		LoBo	LoBo		LoBo	LoBo	
<i>Batch number</i>															
1	0.003	0.171	0.234	0.013	0.203	0.247	0.006	0.198	0.251	0.006	0.196	0.250	0.006	0.196	0.250
2	0.008	0.302	1.692	0.101	0.483	1.373	0.016	0.301	1.362	0.017	0.302	1.250	0.015	0.401	1.231
3	0.002	0.123	0.241	0.016	0.129	0.160	0.016	0.154	0.152	0.005	0.136	0.160	0.006	0.132	0.158
4	0.003	0.206	0.331	0.011	0.236	0.290	0.009	0.236	0.286	0.009	0.234	0.285	0.009	0.232	0.290
5	0.009	0.276	0.702	0.050	0.335	0.752	0.048	0.300	0.743	0.048	0.309	0.728	0.048	0.318	0.738
6	0.001	0.074	0.100	0.006	0.081	0.101	0.006	0.093	0.106	0.004	0.087	0.113	0.005	0.084	0.110
7	0.001	0.125	0.150	0.002	0.133	0.160	0.002	0.142	0.160	0.002	0.136	0.159	0.002	0.135	0.158
8	0.007	0.287	0.482	0.017	0.320	0.452	0.016	0.344	0.454	0.018	0.336	0.483	0.019	0.323	0.477
9	0.001	0.053	0.102	0.003	0.107	0.164	0.003	0.118	0.165	0.003	0.109	0.172	0.003	0.110	0.171
10	0.001	0.039	0.045	0.004	0.043	0.051	0.012	0.049	0.051	0.003	0.045	0.051	0.003	0.045	0.050
11	0.003	0.623	0.705	0.143	0.620	0.723	0.143	0.623	0.718	0.010	0.620	0.723	0.008	0.620	0.720
<i>Mean</i>	0.003	0.207	0.435	0.033	0.244	0.407	0.025	0.232	0.404	0.011	0.228	0.398	0.011	0.235	0.396

Table V: Model error percentages per batch and methodology.