Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Ingegneria Industriale

Ciclo XXIX

Multivariate techniques applied on spectroscopic data for process analysis and monitoring

Settore scientifico disciplinare di afferenza

ING-IND/26

Presentata da Alessandra Taris

Coordinatore Dottorato Prof. Francesco Aymerich

Tutor Massimiliano Grosso, PhD

Esame finale anno accademico 2015 – 2016

Tesi discussa nella sessione d'esame marzo – aprile 2017

Give your heart to what really means

To my second family:
friends

Contents

Abstract

Process analysis and monitoring has become essential in industry to ensure improvement of the process performances and to maintain a specific product quality. To this aim, spectroscopy represents an innovative tool that allows to overcome the issues encountered with conventional analytical techniques (e.g. gas chromatography), since it is fast and non-destructive and can give information about the chemical state of the process in real time. Nevertheless, due to the huge amount of information present in the collected data, the interpretation and information extraction is not a straightforward task. For this purpose, multivariate techniques significantly aid the treatment of the data and allow to infer information about the system analyzed.

In this thesis, four systems are investigated by means of spectroscopy to show the variety of problems that may arise when dealing with complex and highly informative data coming from different spectroscopic techniques. To this aim, different multivariate techniques are explored and their potentialities and limitations are shown: (i) Strategies based on Principal Component Analysis and Partial Least Squares Regression are suggested for an improved and more robust quality monitoring of liquid commercial detergents; (ii) Moving Window Principal Component Analysis is proposed for the monitoring of an evolving process like the crystallization of an Active Pharmaceutical Ingredient in order to detect the nucleation; (iii) Time Window Statistical Total Correlation Spectroscopy combined with Multivariate Curve Resolution are proposed to investigate the setting reaction of a cementing material; (iv) Multivariate Curve Resolution is employed to infer information from hyperspectral data about the dissolution of a surfactants paste.

Therefore, multivariate techniques applied to spectroscopic data demonstrate capable of achieving the following results:

a) in case of commercial detergents, they correctly classify observations that do not agree with the reference conditions. Moreover, the approach proposed is able to assess when the

estimation of the compounds concentration cannot be considered accurate, this scenario may occur when the deviations of one compound is not taken into account during model calibration;

b) for the crystallization of the pharmaceutical ingredient, the nucleation is accurately detected;

c) spectra and concentration of the compounds involved in the setting reaction of a cementing material are estimated and time evolution of the process can be tracked;

d) the dissolution rate of the surfactants present in the paste is estimated.

As a result, multivariate methods applied to spectroscopic data reveal essential to treat data and aid process understanding and monitoring.

Sommario

L'analisi e il monitoraggio di processo sono diventati di fondamentale importanza per garantire le prestazioni del processo e mantenere la qualità del prodotto. A tal scopo, la spettroscopia rappresenta uno strumento innovativo che permette di superare le problematiche che si incontrano con le tecniche analitiche convenzionali (per esempio, la gas cromatografia), poichè è veloce e non distruttiva e può fornire informazioni sullo stato chimico del processo in tempo reale. Tuttavia, a causa della grande quantità di informazioni presenti nelle misure raccolte, l'interpretazione e l'estrazione di informazione non è un compito semplice. A tal proposito, le tecniche multivariate agevolano significativamente il trattamento dei dati e permettono di inferire informazioni sul sistema analizzato.

In questa tesi, quattro sistemi sono indagati mediante misure spettroscopiche per mostrare la varietà di problemi che possono sorgere quando si trattano dati complessi e altamente informativi provenienti da differenti tecniche spettroscopiche. Per questo motivo, sono state esplorate differenti tecniche multivariate e sono mostrate le loro potenzialità e limitazioni: (i) si suggeriscono strategie basate sulla Principal Component Analysis e Partial Least Squares Regression per un migliore e più robusto monitoraggio di qualità dei detergenti commerciali liquidi; (ii) la Moving Window Principal Component Analysis è proposta per il monitoraggio di processi che si evolvono come la cristallizzazione di un Ingrediente Farmaceutico Attivo per identificare la nucleazione; (iii) la Time Window Statistical Total Correlation Spectroscopy insieme alla Multivariate Curve Resolution sono proposte per indagare la reazione di formazione di un materiale cementizio; (iv) la Multivariate Curve Resolution è utilizzata per ottenere informazioni sulla dissoluzione nello spazio e nel tempo di una pasta costituita da tensioattivi a partire da dati iperspettrali .

Perciò, le tecniche multivariate applicate a dati spettroscopici si dimostrano capaci di raggiungere i seguenti risultati:

a) Nel caso di detergenti commerciali, le osservazioni che non rispecchiano le condizioni di riferimento sono classificate correttamente. Inoltre, l'approccio proposto identifica quando la

stima della concentrazione dei composti non può essere considerata accurata;

b) Riguardo la cristallizzazione dell'ingrediente farmaceutico, la nucleazione è stata individuata in modo accurato;
c) Gli spettri e la concentrazione dei composti coinvolti nella reazione di presa di un materiale cementizio sono stati stimati e l'evoluzione temporale del processo può essere seguita;
d) La velocità di dissoluzione dei tensioattivi presenti nella pasta è stata valutata.

Di conseguenza, i metodi multivariati implementati su misure spettroscopiche si rivelano essenziali per trattare i dati e agevolare la comprensione e il monitoraggio di processo.

Chapter 1

Introduction

This introductory Chapter illustrates the motivations that lead to the development of this Thesis. The outline of the Thesis is also presented, describing the content of each Chapter. Finally, conference and journal papers derived from the present work are listed.

1.1 Motivations

The improvement of the process performances and the fulfillment of specific product quality have become relevant issues to address in industry nowadays. To this aim, process analysis and monitoring can ensure the detection of deviations from the reference conditions, the identification of the corresponding cause during a process and a clear understanding of the process occurring. Particularly, product quality represents a key factor for the manufacturer to maintain competitiveness in the market. The purpose is not only to satisfy the customer needs, but also to comply to safety and environmental regulations at a minimum cost. The conventional approach to assess product quality in industry consists of collecting samples at the completion of the process and testing them in a Quality Control laboratory with off-line conventional analytical techniques (e.g. chromatography) that can be destructive and time consuming. Although this procedure ensures a high level of product quality, in case the product does not meet certain specifications, it has to be reprocessed, implying an increase of cost, time waste and perhaps losses of high-value products. On the other hand, on-line monitoring of the quality attributes during production process would allow a prompt intervention and improvement of the process performances and maintenance. This leads to reduction of product waste, improvement to product quality, reduction of byproducts, energy saving, that finally results in cost reduction. Traditionally, online process monitoring is carried out by monitoring process variables like temperature, pressure, that does not always ensure to meet the target quality of the product. On the other hand, conventional analytical techniques are not always a feasible solution for on-line quality monitoring, since they introduce time delay in the control system response.

To this aim, spectroscopy represents an innovative, fast and non-invasive analytical technique that give chemical information about the sample being analyzed and its composition (Kourti, 2006). It can be used to estimate compounds concentration in a product, for process understanding, transient processes monitoring and to infer information of the kinetic and dynamics of the phenomena investigated, it is suitable for on-line monitoring of process and product quality. Indeed, it provides

information about the chemical state of a process, that cannot be always obtained through process variables like temperature, pressure etc. The advantage is the significantly reduced sampling time and no need of physical sampling of the process (Gurden et *al.*, 2002).
Nevertheless, since hundreds of spectra implying thousands of spectral variables can be measured during an experiment (or during a process), the interpretation and the extraction of useful information from collected data through traditional univariate methods is not always straightforward. In the following, the principles underlying these methods and the possible drawbacks will be illustrated.

Typical techniques such as inverse calibration and peak integration are the most common procedures employed to estimate the concentration of compounds present in the sample from spectroscopic data (Brereton, 2000). They consist of monitoring one or a finite number of intensities that are characteristic of the compounds investigated (or integrating area under peaks of interest).
However, these approaches cannot be always employed because it implies that the main compounds are all known *a-priori*. In fact, tens or hundreds of unknowns may be present in samples, whereas sometimes our interest is mainly focused on the quantification of a few of them. On the other hand, it is not always feasible to design samples (standards) for all the potential components present in real samples. Moreover, since univariate methods imply that each peak (intensity) corresponds to only one component, their implementation is not possible when peak overlap occurs, and this issue may lead to serious estimation errors. As a final remark, the integration of single peaks in a manual fashion, can be a very time consuming task.
Furthermore, when an evolving system is investigated, spectra can reflect the phenomena dynamics (e.g. chemical reaction evolution). In order to extract kinetic information, univariate methods are frequently used but the number of the components involved is not always known in advance. For example, unknown intermediate and final species could form. In particular, detection of intermediate species implies the inspection of all the spectra collected during the experiment. This task might reveal to be very time consuming and not always feasible.

The integration of the process (physical state) and spectroscopic (chemical state) measurements can improve the performances of process monitoring (Wong et *al.*, 2008). Traditionally, in order to assess whether the process is under control, univariate classical control charts such as the Shewhart chart (Shewhart, 1931), the CUSUM plot (Woodward & Goldsmith, 1964) and the EWMA chart (Hunter, 1986) are employed. Process variables (temperature, pressure) and their control limits are reported. However, when the number of variables to monitor is relevant (hundreds or even thousands), this approach could lead to hundreds or more control charts, that are not easy to manage. In addition, when few underlying events are driving a process, all the process variables are simply different manifestations of these events and they may be highly correlated and univariate control charts can give misleading outcomes, since they do not take into account the possible correlation between variables (MacGregor & Kourti, 1995).
Therefore, in order to ensure an effective process monitoring and to correctly treat spectral data, proper mathematical methods are required. Multivariate techniques represent a powerful tool to reveal hidden or relevant information when dealing with highly informative and complex data. The use of multivariate statistical methods for the analysis of analytical data is also indicated as chemometrics (Wold, 1995). They exploit the correlation between variables and are able to find patterns and structures among the data not otherwise possible. Hence, they allow to overcome the issues that one may encounter with univariate method (control charts, peak tracking or integration), since they are capable of detecting deviation from normal behavior also in case of highly correlated variables and are definitely suitable for on-line process monitoring and kinetic modeling. Although they are not exempt from issues, their potentialities allow to explore and combine their features to improve data treatment and overcome their limits.

1.2 Thesis outline

This Ph.D Thesis deals with development of approaches based on multivariate techniques to use for spectroscopic data treatment and improve process analysis and monitoring. Particularly, different four

systems are considered such as liquid commercial detergents, an Active Pharmaceutical Ingredient, a dish paste and a cementing material. The variety of the systems represents an overview of the possible issues that may arise during data treatment and it shows how the multivariate techniques can flexible and useful tools to extract information from spectroscopic data and overcome the problems met with the Univariate Technique approaches.
Table 1 summarizes the case studies investigated and the spectroscopic techniques employed in this thesis.
A summary of this Thesis is shown in the following list, where a brief description of each Chapter is given.

Chapter 2 describes in detail the multivariate techniques typically used for process analysis and monitoring.

Chapter 3 gives an overview of the main contributions of this thesis to the state of art.

Chapter 4 deals with the on-line monitoring of commercial liquid detergent mass production.

Chapter 5 illustrates the approaches proposed for the in-line monitoring of cooling crystallization of an Active Pharmaceutical Ingredient (Isonicotinamide).

Chapter 6 investigates the dissolution of a dish paste.

Chapter 7 is focused on methods to help the understanding of the setting reaction of a ceramic material.

Chapter 8 summarizes the main conclusions of the present work.

System investigated	Aim	Spectroscopic technique
Commercial detergents	Quality control	Infrared
Active Pharmaceutical Ingredient	Crystallization monitoring	In situ Infrared
Cementing material	Reaction monitoring	In situ X-ray Powder Diffraction
Surfactants paste	Dissolution investigation	Confocal Raman Microscopy

Table 1 – Overview of the case studies investigated and the spectroscopic techniques employed in this thesis.

1.3 Conference Papers and publications in Journals

Some of the work present in this Thesis has been presented in international congresses and published or submitted in international journal papers.

Conference papers

Taris A., Grosso M., Zonfrilli F., Guida V. (2015). Quality control of industrial detergents through infra-red spectroscopy measurements coupled with partial least square regression. Chemical Engineering Transactions, 43, 1549-1554. ICheaP-12 International Conference, Milan (Italy).

Taris A., Grosso M., Viani A., Brundu M., Guida V. (2015). Reaction monitoring of cementing materials through multivariate techniques applied to time-resolved synchrotron X-ray diffraction data. Chemical Engineering Transactions, 43, 895-900. ICheaP-12 International Conference, Milan (Italy).

Taris A., Grosso M., Brundu M., Guida V., Viani A. (2015). Reaction Monitoring of Cementing Materials through Multivariate Techniques Applied to In Situ Synchrotron X-Ray Diffraction Data. Computer Aided Chemical Engineering, 37, 1535-1540. ESCAPE25, Copenhagen (Denmark).

Taris A., Grosso M., Zonfrilli F., Guida V. (2014). Statistical control of commercial detergents production through Fourier Transform Infra-Red spectroscopy. Computer Aided Chemical Engineering, 33, 601-606. ESCAPE24, Budapest (Hungary).

Taris A., Grosso M., Viani A., Brundu M., Guida V. (2015). Combined Multivariate Techniques for Improved Reaction Monitoring Applied to In Situ X-Ray Diffraction Data. Scandinavian Symposium on Chemometrics, Pula (Italy).

Journal papers

Hansen T. B., Taris A., Rong B-G, Grosso M., Qu H. (2016). Polymorphic behavior of isonicotinamide in cooling crystallization from various solvents. Journal of Crystal Growth, 450, 81–90.

Taris A., Grosso M., Brundu M., Guida V., Viani A. Application of combined multivariate techniques for the description of time-resolved powder X-ray diffraction data. Accepted for publication in Journal of Applied Crystallography.

Taris A., Hansen T. B., Rong B-G, Grosso M., Qu H. Statistical process monitoring of cooling crystallization through Moving Window PCA and contribution plots applied to in situ infrared data. Submitted to Organic Process Research & Development.

Chapter 2

Multivariate techniques - State of art

This Chapter describes the multivariate techniques usually employed to analyze and extract information from multivariate data for process analysis and monitoring.

Multivariate techniques can be used when several measurements (variables) are collected for each sample or observation. The linear ones exploit the linear relationship between variables and can be employed for on-line process monitoring and kinetic modeling.

2.1 Data representation

The experimental data can be arranged in a matrix $\mathbf{X}_{(I\times J)}$, where I are the number of observations and the *i*-th row (i = 1, … I) of the matrix represents the experimental spectrum $\mathbf{x}_{(1\times J)}$ collected along the different *J* spectral variables (wavenumbers, diffraction angles 2θ, etc.), as shown in Equation (1). The element x_{ij} is the value of the *j*-th variable measured for *i*-th observation (i=1,…, I; j =1,…, J).

$$\underset{(I\times J)}{\mathbf{X}} = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{I1} & x_{I2} & \cdots & x_{IJ} \end{vmatrix} \quad (1)$$

2.2 Data pre-processing

The experimental data matrix is usually pre-processed before the implementation of multivariate techniques (Zeaiter & Rutledge, 2009).

2.2.1 Baseline correction and smoothing

Since the experimental measurements can be noisy and the baseline can drift during the experiment, smoothing and baseline correction of spectra can be required. Concerning the smoothing, generally Savitzky–Golay filter are used (Savitzky & Golay, 1964), where a window is moved along the spectrum along the spectral variables direction and spectrum is approximated with a polynomial function of *n*-order. Regarding the baseline correction, the baseline can be approximated with a *n*-order polynomial curve obtained through Savitzky–Golay filter (Savitzky & Golay, 1964), other approaches used are linear interactive baseline

correction algorithm (e.g., Hrovat, 2009) or minimized cost function (Mazet et *al.*, 2005).

2.2.2 Variables transformation

Data are pre-processed since spectral variables can have different order of magnitude (McGregor, 1995; Romagnoli & Palazoglou, 2012) or the measurements might not be normally distributed. These features can affect the multivariate analysis.

Mean centering

The mean centering is the most common method, where each element x_{ij} are centered according the Equations (2) and (3).

$$\mathrm{x}_{ij} = \mathrm{x}_{ij} - \bar{\mathrm{x}}_j \tag{2}$$

where, in Equation (2)

$$\bar{\mathrm{x}}_j = \frac{\sum_{i=1}^{I} \mathrm{x}_{ij}}{I} \tag{3}$$

Unity variance

Since multivariate techniques can be scale dependent, the unity variance (UV) standardization is usually employed and this transformation is reported in Equation (4).

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{4}$$

Where $s_j = \sqrt{\frac{\sum_{i=1}^{I} (x_{ij} - \bar{x}_j)^2}{I-1}}$ is the standard deviation of the *j*-th variable.

This transformation aims to reduce the influence of the scale on the variables and make them comparable, such that the mean of the transformed *j*-th variable is zero and its standard deviation is unity.

2.3 Multivariate analysis

Linear Multivariate Statistical Techniques have revealed as a powerful tool to extract chemical information from multicomponent spectra. The main premise of spectroscopic data is the linearity, assuming that the interaction between components are linear and light scattering does not occur. Indeed, each observed spectrum can be seen as a linear combination of the spectra of each species in the sample. The experimental matrix $\mathbf{X}_{(I\times J)}$ can be thus decomposed as expressed in Equation (5).

$$\underset{(I\times J)}{\boldsymbol{X}} = \underset{(I\times A)}{\boldsymbol{T_A}} \cdot \underset{(A\times J)}{\boldsymbol{P_A^T}} + \underset{(I\times J)}{\boldsymbol{E}} = \underset{(I\times J)}{\widehat{\boldsymbol{X}}_{\boldsymbol{A}}} + \underset{(I\times J)}{\boldsymbol{E}} \tag{5}$$

In Equation (5), A is the number of pseudo-components, while $\mathbf{T}_A$ can be seen as the pseudo-concentration (or abundance) matrix whose columns, $\mathbf{t}_{a(I\times 1)}$ represent the pseudo-concentration (or abundance) of the a-th component estimated for the i-th observation, whereas each column of the $\mathbf{P}_A$ matrix, $\mathbf{p}_{a(J\times 1)}$ represents the estimated pseudo-spectrum of the a-th component. The matrix $\mathbf{E}$ is the residual matrix that represents the information not captured by the decomposition. The superscript T denotes the matrix transpose. $\widehat{\boldsymbol{X}}_{\boldsymbol{A}}$ is the estimation of the experimental data matrix. Depending on the method used, $\mathbf{T}_A$ and $\mathbf{P}_A$ matrices have different definition. The most popular approach is Principal Component Analysis (Joliffe, 2002), others employed are non-iterative such as the Evolving Factor Analysis (Maeder, 1987), Sub-window Factor Analysis (Manne et *al.*, 1999) or iterative, requiring an initial estimate for the matrix $\mathbf{T}_A$ and/or $\mathbf{P}_A$, such as the Multivariate Curve Resolution-Alternative Least Squares (Tauler, 1995).

2.4 Static Principal Component Analysis

The Principal Component Analysis (PCA) is a well-known multivariate technique (Joliffe, 2002) used to compress data and extract relevant information from experimental data, it is indicated as static PCA, when

applied to stationary data. It is based on the decomposition of the generic data matrix $\mathbf{X}_{(N\times J)}$, as reported in Equation (5).

$$\underset{(N\times J)}{\boldsymbol{X}} = \underset{(N\times J)}{\boldsymbol{T}} \cdot \underset{(J\times J)}{\boldsymbol{P^T}} \tag{6}$$

Matrix **P** is computed as the eigenvector matrix of the covariance matrix **S**. This is semipositive and defined in Equation (7).

$$\underset{(J\times J)}{\boldsymbol{S}} = \frac{\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}}{I-1} \tag{7}$$

The loading matrix $\mathbf{P}^{\mathrm{T}}$ is orthonormal, which implies that $\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{-1}$. It is defined as the rotation matrix, since it is possible to rotate the original space into the new space. Each column of the matrix $\mathbf{P}^{\mathrm{T}}$ describes the relationship between the spectral variables and the *j*-th principal component.
On the other hand, **T** is defined as the score matrix that represents the projections of the spectral variables onto the new subspace identified by the principal components. Each score (i.e., each column $\mathbf{t}_j$ of **T**) is orthogonal ($\mathbf{t}_j^T \cdot \mathbf{t}_k = 0\ \forall j,k$).
It can be noted that the *j*-th eigenvalue λ_j in Equation (8) represents the variance explained by the *j*-th principal component. This value depends on the relative importance of the *j*-th principal component, in other words its ability to capture the variability of the data.

$$\lambda_j = \frac{(\mathbf{T}^{\mathrm{T}}\cdot\mathbf{T})_{jj}}{I} = \frac{1}{I}\sum_{i=1}^{I} t_{ji}^2 \tag{8}$$

The algorithm used for the estimation of the matrices **P** and **T** are iterative such as the Singular Value Decomposition (SVD) (Anderson et *al.*, 1999) or the Nonlinear Iterative Partial Least Squares (NIPALS) (Wold, 1975).

2.4.1 PCA model

The aim of the PCA is the approximation of the data matrix **X** such that only relevant information is retained and the first *A* principal component

are considered to describe the variance of the data as reported in Equation (9).

$$\underset{(I\times J)}{\widehat{\boldsymbol{X}}_{\boldsymbol{A}}} = \underset{(I\times A)}{\boldsymbol{T}_{\boldsymbol{A}}} \cdot \underset{(A\times J)}{\boldsymbol{P}_{\boldsymbol{A}}^{\boldsymbol{T}}} \tag{9}$$

Where the matrix $\mathbf{T}_A$ is the projection of the original variables onto the new subspace defined by the first A PCs. $\widehat{\boldsymbol{X}}_{\boldsymbol{A}}$ is the prediction matrix obtained by retaining the first A principal components (PCs).
A key factor in the development of PCA model is the choice of the number of PCs to retain. There are different criteria in literature, but the most used is based on the cumulative variance explained (Joliffe, 2002) as expressed in Equation (10).

$$Var(a) = \sum_{a=1}^{J} \frac{\lambda_a}{\sum_{j=1}^{J} \lambda_j} \times 100\% \qquad \text{with } a=1, 2, \ldots, J \tag{10}$$

Therefore, the implementation of PCA to treat the data matrix **X** involves the following steps:

i. *Pre-processing*

Pre-processing of data matrix as described in Section 2.2.

ii. *Calibration*

the model is built based on the training set, $\mathbf{X}^{\mathbf{c}}_{(\mathrm{N}\times\mathrm{J})}$, to this aim it should be chosen such that it is a quite fair representation of the normal operating condition (NOC). Subsequently, samples mean and covariance matrix, loading matrix **P** are computed. The fundamental step in the PCA modelling is the choice of the number of principal components (the cumulative variance criterion can be employed).

iii. *Prediction*

New observations are projected onto the PCA model to evaluate whether they are consistent with it and they behave according to the training set. Thus, a new multivariate observation $\mathbf{x}_k$ is projected onto the space spanned by the first A principal components through the relationship (11).

$$\mathbf{t}_k = \mathbf{x}_k \cdot \mathbf{P}_A \tag{11}$$

2.5 On-line process monitoring

In order to detect abnormal process behaviour or classify a new observation as belonging to the training set, two statistics, T^2 and Q (or Square Prediction Error, SPE), are employed for statistical process control (MacGregor and Kourti, 1995).

The T^2 is based on the work of Hotelling (1947), for a new observation $\mathbf{x}_k$, the statistic is evaluated through Equation (12).

$$T_k^2 = \mathbf{t}_k^T \cdot \Lambda^{-1} \cdot \mathbf{t}_k \tag{12}$$

where, in Equation (12), $\mathbf{\Lambda}$ is the covariance matrix of the $\mathbf{t}_k$ scores. It is a diagonal matrix whose elements are the eigenvalues λ_j. In practice, the T^2 statistic represents an overall measure of the process variation related to $\mathbf{x}_k$ as it was captured by the PCA model. In other words, it is the distance of the projections from the origin of the new subspace defined by the first A PCs.
Statistical confidence limits for T^2 statistic (Tracy et *al.*, 1992) are calculated by means of Equation (13).

$$T_{lim}^2 = \frac{A \cdot (N^2 - 1)}{N \cdot (N - A)} F_\alpha(A, N - A) \tag{13}$$

Whilst, the Q statistic represents the euclidean distance between the observation and its projection onto the subspace defined by the first A PCs as reported in Equation (14) and describes how well the PCA model predicts the $\mathbf{x}_k$ vector (Jackson, 1991).

$$Q_k = \mathbf{e}_k^T \cdot \mathbf{e}_k, \qquad \mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k = \mathbf{x}_k - \mathbf{t}_k \cdot \mathbf{P}_A \tag{14}$$

where $\boldsymbol{e}_k$ is the difference between the experimental observation $\mathbf{x}_k$ and the value $\hat{\mathbf{x}}_k$ predicted through the PCA model where A PCs are considered.
The upper control limit for Q (see Jackson & Muldholkar, 1979) is defined as:

$$Q_{\lim} = \vartheta_1 \left[1 + \vartheta_2 h_0 \left(\frac{h_0 - 1}{\vartheta_1^2} \right) + \frac{c_\alpha h_0 \sqrt{(2\vartheta_2)}}{\vartheta_1} \right]^{1/h_0} \quad (15)$$

where:

$\vartheta_i = \sum_{j=A+1}^{J} \lambda_j^i$ $\quad$ i=1, 2, 3

$h_0 = 1 - \frac{2\vartheta_1 \vartheta_3}{3\vartheta_2^2}$

c_α is the α-th upper percentile of the standard normal distribution.

Another relationship used to calculate the Q threshold value (Nomikos & MacGregor, 1995) is reported in Equation (16), where the Q statistic follows a Chi-square distribution with g_2 degrees of freedom $\chi^2_{g_2,\alpha}$, α is the significance level (usually 5 %).

$$Q_{x,\lim} = g_1 \cdot \chi^2_{g_2,\alpha} \quad (16)$$

The constants g_1 and g_2 are obtained as reported in Equation (17), where $\bar{Q}$ and s_Q^2 are the mean and the variance of the Q statistics estimated for the observations belonging to the training set.

$$g_1 = \frac{s_Q^2}{2\bar{Q}} \quad \text{and} \quad g_2 = 2\bar{Q}^2 \quad (17)$$

2.5.1 Geometric interpretation of T^2 e Q

In order to correctly interpret the results obtained through T^2 e Q statistics, a key aspect to consider is the different role that they have during process monitoring. Indeed, they measure different deviations from nominal behavior (Qin, 2003; Wise & Gallagher, 1996). In Figure 1 an illustrative example is reported, where a bidimensional space (x_1-x_2) is considered. The training set is depicted as grey circles and data are highly correlated. Thus, the direction (dashed black line) where the data (grey circles) lie represents the first principal component (PC_1) that identifies the new subspace. Therefore, each point is projected onto the component and O′ is

the multivariate mean of these projections. Two new samples (depicted as red and green squares) show the different role of the statistics. Their coordinates in the new subspace are the projections $\hat{C}$ e $\hat{D}$, respectively.

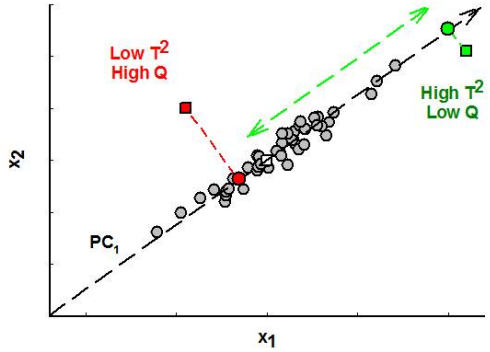

Figure 1 – Illustrative example to show the different role of T^2 e Q statistics. The green square is classified as out-of-control by T^2 statistic, since its projection onto the new subspace is distant from the new origin O′. The red square is out-of-control according to Q statistic since the Euclidean distance from its projection is higher than the values calculated for the training set.

As previously explained, T^2 is the distance of the projections from the origin of the new subspace defined by PC_1, thus it measures a deviation within the subspace. This distance is represented by the segment $\hat{C}O'$ and $\hat{D}O'$ for the red and green square, respectively. As a result, the sample D is classified out-of-control since the distance from the mean O′ is high. On the other hand, the Q statistic is the Euclidean distance of the points C and D from their projections onto the first component $(\hat{C}$ e $\hat{D})$. Thus, the red square is identified as out-of-control, since it moved off the plane (Wise & Gallagher, 1996). Hence, the role of these statistics is asymmetric: T^2 is able to detect only significant deviations from the multivariate mean (green square), in this case the correlation is still valid but the normal behavior is shifting away from the mean along the same direction identified by the first component, this scenario may not be necessarily a real fault. Regarding Q statistic, an abnormal situation is detected when it breaks the normal process correlation (red square) and residuals become higher, thus the PCA model cannot describe the new data. It should be noted that T^2 is associated to the subspace defined by the components and they are characterized by large variations, while Q is related to the residual subspace that contains mainly noise. As a consequence, the

normal region defined by the control limit for T^2 is usually much larger than that of Q. Therefore, the fault should be more relevant to exceed the T^2 control limit rather than the Q one (Qin, 2003). For this reason, in most cases Q is preferable than T^2 for fault detection.

2.5.2 T^2 and Q contribution plot

Contribution plot is used to identify which variables are most contributing to T^2 and Q statistics and then are deviating from the typical behavior (Alcala & Qin, 2011; Westerhuis et *al.*, 2000). The contribution index $\mathbf{c}_k^Q$ of the Q statistics can be evaluated for the k-th observation through Equation (18). Each element of the vector corresponds to the ratio between the contribution of each variable to Q statistics, $\mathbf{e}_k^2$, and the expected value (Alcala & Qin, 2011).

$$\mathbf{c}_k^Q = \frac{\mathbf{e}_k^2}{diag(\mathbf{S}\widetilde{\mathbf{P}}_A\widetilde{\mathbf{P}}_A^T)} \tag{18}$$

Where $\widetilde{\mathbf{P}}_A = \mathbf{I} - \mathbf{P}_A$ is the loading matrix in the residual space.
Similar definition for the contribution vector $\mathbf{c}_k^{T^2}$ of T^2 is given in Equation (19), where $\mathbf{\Lambda}$ is the eigenvalues matrix.

$$\mathbf{c}_k^{T^2} = \frac{\left(\mathbf{P}_A\mathbf{\Lambda}^{-0.5}\mathbf{P}_A^T\mathbf{x}_k\right)^2}{diag(\mathbf{S}\mathbf{P}_A\mathbf{\Lambda}^{-1}\mathbf{P}_A^T)} \tag{19}$$

The limit value for the contributions $\mathbf{c}_k^Q$ and $\mathbf{c}_k^{T^2}$ is generally given by a $\chi^2_{1,\alpha}$. Nevertheless, when the multiple statistical test is carried out on a multivariate sample, the probability of running into false positive values increases with J variables. To avoid excessive false positives and then reduce the chance of a type I error, the Bonferroni correction can be employed (Broadhurst & Kell, 2006; Armstrong, 2014), since it reduces the significance level for the single variable from α to α/J.

2.5.3 On-line process monitoring procedure

On-line process monitoring is implemented following the subsequent steps: first, a PCA model, based on the reference data of the process, $\mathbf{X}^{\mathbf{c}}_{(I\times J)}$, has to be identified. Therefore, the loading matrix **P**, are determined and the number of principal components is selected. In addition, T^2 and Q limits are evaluated according to Equations (13) and (15). For on-line monitoring, the new data points $\mathbf{x}_k$ are projected onto the model space defined by the PCA model, spanned by the retained A loading vectors as reported in Equation (11). Then, the associated values of the T^2 and Q statistics are calculated by means of Equations (12) and (14) and usually reported in control charts. The occurrence $Q_k > Q_{lim}$ and/or $T^2_k > T^2_{lim}$ may be indicative of abnormal process behavior or the observation is not consistent with the reference ones (MacGregor and Kourti, 1995; Romagnoli & Palazoglou, 2012). Once out-of-control observations are detected, contribution plots are usually employed, in order to identify which j-th spectral variables are significantly deviating from the reference behavior and are contributing to T^2 and Q statistics.

2.6 Partial Least Squares

PLS represents a powerful multivariate statistical tool for the quantitative analysis of spectroscopic data that enables to overcome problems common to this data such as collinearity, peak overlaps and interactions and it can be seen as a feasible method for multivariate calibration. It can be also employed for process-monitoring (Godoy et *al.*, 2014), but is also able to determine compounds concentration (PLS-Regression) (Wold et *al.*, 2001, Meng et *al.*, 2014), and for pattern recognition (PLS-Discriminant Analysis) (e.g., Szymańska et *al.*, 2012).

2.6.1 PLSR- Model calibration

Given a predictor matrix $\mathbf{X}_{(I\times J)}$ and a response matrix $\mathbf{Y}_{(I\times N)}$ the PLS algorithm projects **X** and **Y** onto a low-dimensional space defined by a small number of latent variables A (Li et *al.*, 2010) as expressed in Equation (20) and (21). Similarly to PCA, the choice of the number of

latent variables ($A < J$) is a key factor to describe adequately data, the cumulative variance criterion (Equation (10)) can be employed for PLS matrices **X** and **Y** as well.

$$\underset{\mathrm{I\times J}}{\mathbf{X}} = \underset{\mathrm{I\times A}}{\mathbf{T_A}} \cdot \underset{A\times J}{\mathbf{P_A^T}} + \underset{\mathrm{I\times J}}{\mathbf{E}} \tag{20}$$

$$\underset{\mathrm{I\times N}}{\mathbf{Y}} = \underset{\mathrm{I\times A}}{\mathbf{T_A}} \cdot \underset{A\times N}{\mathbf{Q^T}} + \underset{\mathrm{I\times N}}{\mathbf{F}} \tag{21}$$

Where $\mathbf{T_A}$ is the orthonormal score matrix, $\mathbf{P_A}$ and **Q** are the loading matrices for **X** and **Y**, respectively. **E** and **F** are the residuals matrices of **X** and **Y**. In general, **X** and **Y** can be pre-processed and scaled to unity variance and mean centred. The basic idea in PLS-R is that the covariance between **X** and **Y** should be maximized and there are several ways to solve the maximum optimization problem and compute PLS model matrices **P** and **Q**. In this thesis, the SIMPLS algorithm developed by De Jong (1993) was used since it appears faster and easier to interpret than nonlinear iterative partial least-squares one (NIPALS). PLS can be implemented to infer a single response variable (PLS1) or multiple response variables (PLS2). Here, PLS2 was adopted as it seemed more appropriate for process monitoring. This sounds reasonable since the joint regression of multiple response variables should provide more information than the ones collected by building *N* different independent PLS models (Li et *al.*, 2010).

2.6.2 PLS-R: quality variables prediction

PLS method can predict the *k*-th sample concentration $\mathbf{y}_k$ from the corresponding spectrum $\mathbf{x}_k$ (MacGregor et *al.*, 1994) as expressed in (22) and (23).

$$\underset{1\times N}{\hat{\mathbf{y}}_{\boldsymbol{k}}} = \underset{1\times J}{\mathbf{x_k}} \cdot \underset{J\times N}{\mathbf{B}} \tag{22}$$

Where

$$\underset{\mathrm{J\times N}}{\mathbf{B}} = \underset{\mathrm{J\times A}}{\mathbf{R}} \cdot \underset{A\times N}{\mathbf{Q}^{\mathrm{T}}} \tag{23}$$

In Equations (22) and (23) **R** is the pseudo-inverse of the $\mathbf{P_A}$ matrix and **B** is the regression coefficients matrix estimated through the matrices **R** and $\mathbf{Q}^T$. The prediction ability of PLS-R for the training set can be evaluated through the root mean square error of calibration (RMSEC), and for the prediction set through the root mean square error of prediction (RMSEP). They are determined according to Equation (24) and (25), where *C* and *V* are the number of samples in the training set and in the prediction set, respectively.

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{C}(y_i - \hat{y}_i)^2}{C}} \tag{24}$$

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{V}(y_i - \hat{y}_i)^2}{V}} \tag{25}$$

2.7 PCA-based approaches for transient data

When data depend on an external variable *t* (time, temperature, pressure, etc.), the *I* observations are collected over the time and/or with the temperature, this implies that data contain dynamic information. Although Static PCA has been flexibly applied to spectroscopic data to reveal deviations or transitions from the reference conditions (Alcalà et *al.*, 2010; Kogermann et *al.*, 2004, Lin et *al.*, 2006), when data contain dynamic information and are dependent on external variable *t* (time, temperature, pressure, etc.), the correct detection of faults cannot be always guaranteed during on-line monitoring. As a result, misclassification of new observations may occur, i.e., it can detect excessive false alarm. In fact, in a dynamic system data do not respect the assumption of independence with the variable *t* (Ku et *al.*, 1995) and, actually, spectral variables and related statistics, such mean and covariance, are changing with respect to *t.* Therefore, since the training set remains static, it is not able to represent the current status of the process. In fact, during a transient process, variables can intrinsically change (increase or decrease), although this occurrence does not necessarily imply that the system is out-of-control. Therefore, when future observations are projected onto a static PCA, they

could be misidentified as deviation from the reference conditions (and thus classified as out-of-control).

Different methodologies have been developed to monitor transient processes, such as dynamic PCA (DPCA) for auto-correlated data (Ku et *al.*, 1995; Lu et *al.*, 2005), while Recursive PCA (Li et *al.*, 2000) and Moving Window PCA (Jeng, 2010; Wang et *al.*, 2005) for non-stationary data. An interesting comparison of these three methods and works presented so far can be found in De Ketelaere et *al.* (2015). DPCA is the combination of PCA with Autoregressive Integrated Moving Average (ARIMA) model. The extended data matrix used for model calibration is composed of additional time shifted replicates of the original variables. Although it has been successfully applied, Kruger et *al.* (2004) show that the scores of DPCA will be inevitably autocorrelated. This leads to an increased rate of false alarms through T^2. Whereas, Q statistic seems not affected by autocorrelation of the scores. Nevertheless, it is vulnerable to nonstationarity for the same reason as static PCA. Moreover, since DPCA use many more variables than static PCA to build the extended matrix, the interpretation of the contribution plots is more difficult than static PCA. To overcome these shortcomings and limit the influence of older observations on the estimation of the mean and covariance, other methods can be employed such as Recursive PCA (RPCA) and Moving Window PCA (MWPCA).

Useful guidelines to choose the correct parameters in RPCA and MWPCA are provided by Schmitt et *al.* (2016). In detail, these latter methods involve updating the PCA model considering different datasets. RPCA includes new observations and exponentially downweights old ones to evaluate the mean and covariance matrix used in PCA. Hence, a forgetting parameter between 0 and 1 should be selected in order RPCA to give lower weight to older observations and it may be determined through the minimization of the sum of squared prediction errors (SSPE) of the model (Schmitt et *al.*, 2016). During the implementation of RPCA, the model is not updated when an observation is classified as out-of-control. However, sometimes the choice of the forgetting factor requires a priori knowledge of likely fault conditions (Wang et *al.*, 2005). Furthermore, as the dynamic process evolves, older data can become unrepresentative of the varying process. As pointed out by Jeng (2010) and He & Yang (2008) in presence of drifts, the number of false alarms may increase as the data size

becomes larger, since RPCA leads to a slower speed of model adaptation than MWPCA that can overcome some of these deficiencies of RPCA, building a more suitable adaptive process model. As in static PCA, it consists of calibration and prediction steps but on a finite time window. Nevertheless, the training set is not static but it is continuously renewed by removing the oldest observation and adding the newest one as long as the window moves. On the other hand, discarding observations too quickly can lead to too rapidly varying models and ineffective process monitoring. As RPCA, the model is not updated when an observation is detected as out-of-control. Since a fixed size window is moved along the data and a training set, $\mathbf{X}^{\mathbf{c}}_{(\mathrm{L}\times\mathrm{J})}$, is selected, the size of the moving window, L, is a key parameter to choose. It depends on the speed at which the parameters (mean and covariance) change. There are different criterions used to select the proper size of the window. Nevertheless, when the number of available observations is limited, the size L is chosen so that the sum of squared prediction errors (SSPE) of the validation set $\mathbf{X}^{\mathbf{v}}_{(\mathrm{V}\times\mathrm{J})}$ (belonging to the NOC, where V is the size of the validation set) is minimized (for further details see Schmitt et *al.*, 2016; Montgomery, 2008b) as reported in Equation (26).

$$SSPE(l) = \sum_{t=l+1}^{l+V} \left\| \mathbf{x}^{\mathrm{v}}_{\mathrm{t}} - \hat{\boldsymbol{x}}^{\mathrm{v}}_{\mathrm{t,l}} \right\|^2 \tag{26}$$

Where $\hat{\boldsymbol{x}}_{\boldsymbol{t},\mathrm{l}}$ is the observation of the validation set at t (with t=l+1, l+2, …, l+V) predicted through the PCA model built with the observations collected between t=T_0 and t=l (where l=$L_{\min}$, $L_{\min}$+1, …, $L_{\max}$), so that its size is increasing until the maximum.

2.8 Evolving Factor Analysis

When data collected come from an evolving system, local rank exploratory methods are powerful tools to extract information for the resolution of dynamic multicomponent systems, not only the total number of components can be determined, but also the location and evolution of each of these detected components can be inferred (de Juan et *al.*, 2004).
The EFA is a local rank method and aims to monitor the rank evolution of the **X** data matrix (Maeder, 1987). It performs singular values

decomposition on augmented submatrices $\mathbf{X}_i$ (i=1, 2, ...I) in backward and forward time direction as shown in Figure 2.

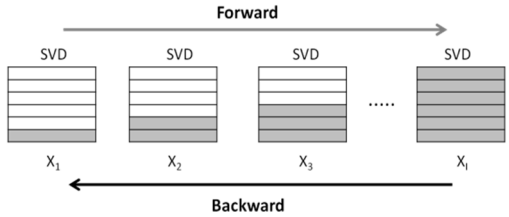

Figure 2 – Graphical representation of the matrix augmentation carried out during EFA implementation.

Combining the eigenvalues evolution obtained during the backward and forward analysis, the EFA plot can be built and information about the number of chemical species involved in the process can be inferred. Estimation of the concentration profiles is eventually carried out. The main advantage of the method is that no a priori knowledge about the chemical species involved in the process is required. Its main drawback is that the mathematical rank of the data matrix does not always correspond to the chemical rank (i.e., the actual number of chemicals in the system), due to the possible occurrence of rank deficiency of the experimental data matrix. In fact, in an evolving system the maximum mathematical rank (Amrhein et *al.*, 1996), i.e. the maximum number of detectable species, is the minimum between the number of independent reactions (N_R) and chemical species in the process (N_C), as reported in Equation (27).

$$rank(\mathbf{D}) = \min\left(N_R + 1, N_C\right) \tag{27}$$

Therefore, EFA may encounter some difficulties to distinguish more reactants (products) which decrease (increase) together in the same direction, and the outcomes of the algorithm could be a linear combination of these compounds.

2.9 Multivariate Curve Resolution

The MCR is a useful tool to investigate complex evolving system. It aims to estimate the spectra of the components and the corresponding concentration profiles in an evolving system (Tauler, 1995). For this

reason, it has been extensively applied to spectroscopic data coming from in situ experiments and hyphenated analytical techniques (De Juan *et al.*, 2014; Garrido et *al.*, 2008; Hantao et *al.*, 2012). The decomposition underlying this multivariate technique $\mathbf{X}=\mathbf{T_A}\cdot\mathbf{P_A}^T + \mathbf{E}$, is the same as that shown in Equation (5), where the matrices $\mathbf{T_A}$ and $\mathbf{P_A}^T$ (generally indicated as **C** and $\mathbf{S}^T$ matrices, respectively) represent the concentration profiles of the pure pseudo-species and spectra of pure pseudo-species, respectively. Unlike PCA, here the constraint of orthogonality between the scores is relaxed and in order that the matrices **T** and **P** have physico-chemical meaning, one could set several constraints as follows (de Juan et *al.*, 2009):

i. Nonnegativity: concentration or response profiles (spectra) are forced to be positive.
ii. Unimodality: it allows concentration or spectra profiles to have a single maximum.
iii. Closure: it forces concentration profiles within the closed system to add up to a certain constant value (the closure constant) to satisfy the mass balance closure condition.
iv. Hard modelling: it shapes concentration profiles and responses according to a mathematical function. In the concentration direction, these functions are typically physico-chemical models (kinetic or equilibrium).

As a consequence, these constraints lead to a more meaningful and interpretable results than PCA.

The MCR algorithm developed by Tauler (1995) and Jaumot *et al.*, (2005) is an Alternating Least Squares iterative optimization procedure that minimizes the norm of the residual matrix **E**. Since it is an iterative method, suitable initial guess $\mathbf{T_A}^0$ for $\mathbf{T_A}$ (or alternatively a feasible $\mathbf{P_A}^0$ for $\mathbf{P_A}$) are required. Particularly, the choice and generation of the starting values is the crucial aspect of the optimization process in order to convey the iteration to chemically meaningful results. The number of compounds in **X** can be determined using PCA or can be known beforehand. If experimental spectra $\mathbf{P}^0$ or concentration profiles $\mathbf{T}^0$ are not available, different methods can be employed, such as EFA and SIMPLe-to-use Interactive Self-modeling mixture Analysis (SIMPLISMA) (Windig, 1997) to estimate the first guesses. The MCR-ALS method consists of

solving alternatively the following two least-squares problems under suitable constraints (Equations (28) and (29)).

$$\boldsymbol{T} \leftarrow \underset{\boldsymbol{T}|\boldsymbol{P}}{min}\|\boldsymbol{X} - \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}}\| \qquad \text{under constraints on } \mathbf{T} \tag{28}$$

$$\boldsymbol{P}^{\boldsymbol{T}} \leftarrow \underset{\boldsymbol{P}|\boldsymbol{T}}{min}\|\boldsymbol{X} - \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}}\| \qquad \text{under constraints on } \mathbf{P}^{\mathrm{T}} \tag{29}$$

In each iteration, $\mathbf{T}$ and $\mathbf{P}^{\mathrm{T}}$ matrices are calculated as shown in Equations (30) and (31).

$$\mathbf{P}^{\mathbf{T}} = \mathbf{T}^{+}\mathbf{X} \tag{30}$$

$$\boldsymbol{T} = (\boldsymbol{P}^{T})^{+}\boldsymbol{X} \tag{31}$$

Where $\mathbf{T}^{+}$ and $(\boldsymbol{P}^{T})^{+}$ are the pseudo-inverse of the $\mathbf{T}$ and $\mathbf{P}^{\mathrm{T}}$. The iterations continue until the convergence criterion is satisfied, i.e. when relative differences in standard deviations of the residuals between experimental data matrix and data estimated through ALS are less than a previously selected value, usually 0.1%.
At the end of the iterations, the percentage of variance explained and the Lack of Fit (Jaumot et *al.*, 2005) can be considered in order to evaluate the goodness of fit. The Lack of Fit is defined as the difference among the input data $\mathbf{X}$ and the data estimated through MCR as reported Equation (32), where x_{ij} is the element of the $\mathbf{X}$ matrix, while I and J are the number of rows and columns of the $\mathbf{X}$ matrix.

$$LOF = \sqrt{\frac{\sum_{i,j}\left(x_{ij} - \hat{x}_{ij}\right)^2}{\sum_{i,j} x_{ij}^2}} \cdot 100\% \tag{32}$$

The percentage of variance explained (Equation (33)) and standard deviation of residuals with to respect experimental data (Equation (34)) are calculated according to the following expressions.

$$R^2_{MCR} = \frac{\sum_{i,j} x_{ij}^2 - \sum_{i,j}\left(x_{ij} - \widehat{x_{ij}}\right)^2}{\sum_{i,j} x_{ij}^2} \cdot 100\% \tag{33}$$

$$\sigma_{MCR} = \sqrt{\frac{\sum_{i,j}\left(x_{ij} - \widehat{x_{ij}}\right)^2}{I \cdot J}} \tag{34}$$

Moreover, the agreement between the spectra (or the concentration) estimated and the experimental ones can be assessed through the Pearson correlation coefficient as reported in Equation (35).

$$r_P = \frac{\sum_{j=1}^{J}\left(y_j - \bar{y}\right)\left(\hat{y}_j - \bar{\hat{y}}_j\right)}{\sqrt{\sum_{i=1}^{I}\left(y_j - \bar{y}\right)^2 \sum_{i=1}^{I}\left(\hat{y}_j - \bar{\hat{y}}_j\right)^2}} \tag{35}$$

Where y_j and $\hat{y}_j$ are the values of the experimental and the estimated spectrum (or concentration) at *j*-th spectral variable (or over the time), and $\bar{\hat{y}}_j$ and $\bar{y}$ are the mean of the experimental and the estimated spectrum (or concentration), respectively.

2.10 Statistical Total Correlation Spectroscopy

The Statistical Total Correlation Spectroscopy, STOCSY, (Cloarec et *al.*, 2005) was firstly introduced by Sâsić et *al.* (2000) based on the same concepts and theory underlying the Two-Dimensional Correlation Spectroscopy (2COS) developed by Noda (1993). It originally aims to aid the detection of potential biomarker molecules in metabonomic studies based on Nuclear Magnetic Resonance spectroscopic data. Since, potentially thousands of different metabolites can be present in complex biosamples, the analysis of the full spectrum and the detection of metabolites and identification of the corresponding spectrum can be quite challenging.

However, the STOCSY is able to overcome this issue, since it takes into account only the synchronous changes of the spectral intensities and exploits the fixed proportionality in a set of NMR spectra between resonances coming from the same molecule.

Considering the case of intensities recorded at different spectral variables, one eventually ends up with a correlation matrix $\mathbf{R}_{(J\times J)}$ whose elements $r_{h,l}$ represent the Pearson correlation coefficient between the intensities at the

spectral variables h and l (h=1, …, J and l=1, …, J) as expressed in Equation (36).

$$r_{h,l} = \frac{\sum_{i=1}^{I}\left(\mathrm{x}_{i,h} - \overline{\mathrm{x}}_h\right)\left(\mathrm{x}_{i,l} - \overline{\mathrm{x}}_l\right)}{\sqrt{\sum_{i=1}^{I}\left(\mathrm{x}_{i,h} - \overline{\mathrm{x}}_h\right)^2 \sum_{i=1}^{I}\left(\mathrm{x}_{i,l} - \overline{\mathrm{x}}_l\right)^2}}$$
$$with \quad h = 1,...,J \qquad l, = 1,...,J \tag{36}$$

where, in Equation (36), the time-averages $\overline{\mathrm{x}}_h = \frac{1}{I}\sum_{i=1}^{I}\mathrm{x}_{i,h}$ and $\overline{\mathrm{x}}_l = \frac{1}{I}\sum_{i=1}^{I}\overline{\mathrm{x}}_{h\,i,l}$ of the h and l intensities are employed. The row vector $\mathbf{r}_{\mathrm{h}}(1\times\mathrm{J})$ of the **R** matrix represents the correlation of the intensity at the h-th spectral variable with the other l intensities (l=1, …, J). When intensities h and l derive from the same chemical species, $r_{\mathrm{h,l}}$ will tend to unity. Thus, two peaks are assumed to belong to the same pattern if r is greater than a correlation cut-off value η.

Therefore, after computing the correlation matrix **R**, the spectrum $\hat{\mathbf{t}}_{a,STOCSY}$ estimated for the a-th species can be obtained by resorting to the following recipe:

1. choose a reference spectrum $\mathbf{x}_{\mathrm{i}(1\times\mathrm{J})}$ from which the pure compound spectra are extracted. In case of evolving systems, it is suggested to use the initial (i.e. i=1) and the final spectrum (i.e. i=I) to estimate the reactants and products spectra, respectively;
2. select the index h^* hereafter referred as the *driver peak*; it should be preferably the highest one (Cloarec *et al.*, 2005);
3. define a *driver* vector $\mathbf{r}_{\mathrm{h}^*(1\times\mathrm{J})}$ whose elements are determined such that
   a. $\mathbf{r}_{h^*,l} = \mathbf{R}_{h^*,l}$ when $\mathbf{R}_{h^*,l} > \eta$ l=1, …, J
   b. $\mathbf{r}_{h^*,l} = 0$ when $\mathbf{R}_{h^*,l} < \eta$
4. the spectrum of the a-th species is eventually computed multiplying the experimental pattern by $\mathbf{r}_{\mathrm{h}}^{\mathrm{d}}$ as reported in Equation (37).

$$\underset{(J\times 1)}{\hat{\mathbf{t}}_{\mathbf{a},STOCSY}} = \underset{(J\times J)}{diag\left(\mathbf{r}_{h^*}\right)} \cdot \underset{(J\times 1)}{\mathbf{x}^{\mathbf{T}}} \tag{37}$$

5. The residual spectrum $\mathbf{x}_e$ is computed by removing the $\hat{\mathbf{t}}_{a,STOCSY}$ from the reference experimental spectrum

$$\mathbf{x}_e = \mathbf{x} - \hat{\mathbf{t}}^T_{a,STOCSY}$$

6. A new driver peak is selected in the residual spectrum and the procedure is iterated A times, until no more significant peaks can be detected.

As widely reported in the literature, the STOCSY technique and its different extensions (see e.g. Robinette et *al.*, 2013) are suited to extract patterns from spectroscopic data. Some problems may however arise when driver peaks belong to more than one compound. This occurrence may lead to a wrong estimation of the spectrum that, in fact, will result as a linear combination of the spectra pertaining more than one involved species. Particularly, in case of evolving systems, STOCSY cannot accurately estimate compounds patterns when the driver peak belongs to both reactants and products.

2.11 Software employed for data treatment

In this thesis, for the implementation of multivariate techniques in-house and preexisting routines in Matlab® R2015a environment are used. For the MCR-ALS algorithm, MCR toolbox for Matlab developed by Jaumot *et al.* (2005) is used.

Chapter 3
Overview of the contributions to the state of art

In this Chapter, a brief overview of the shortcomings that may arise using multivariate techniques is given. The main contributions of this thesis to the state of art are presented as well.

3.1 Procedures for on-line monitoring of continuous and evolving processes

Table 2 summarizes the spectroscopic techniques and multivariate techniques employed in this thesis. In the following the main contributions will be presented.

3.1.1 Continuous processes

Liquid commercial detergents production was considered as case study for on-line monitoring of quality through Infrared Spectroscopy.
The first goal was the detection of out-of-control samples. This is typically carried out resorting to the statistics T^2 and Q (MacGregor and Kourti, 1995). Different methods have been proposed to define a bivariate probability density function of these two statistics (Chen et *al.*, 2004; Qin, 2003). A bivariate probability density function of the statistics T^2 and Q was here suggested and a new operating region was defined (Elliptical Normal Operating Region, ENOR). To this aim, a non linear transformation of these statistics was proposed to allow them to follow a Gaussian dispersion.
The second aim was the estimation of compounds concentration in a robust way. The conventional methods like the inverse calibration and peak integration are not always suitable for on-line monitoring since they show limitations in the quantitative analysis of spectroscopic data. In fact, they it can be time consuming, inaccurate since the spectra can be characterized by collinearity and peak overlaps. On the other hand, Partial Least Squares Regression (PLS-R) enables to overcome these problems (Wold et *al.*, 2001) and can be also implemented for on-line monitoring (Godoy et *al.*, 2014). However, it should be noted that its prediction ability could be worsen if the PLS-R model is calibrated without considering the possible presence of external interferences and may lead to inaccurate results.
Hence, the Q_X statistic was proposed in this thesis to assess the reliability of the prediction in presence of a species whose deviations were not taken into account during model calibration. In presence of out-of-control samples the estimation of the concentration of the other compounds was not considered accurate.

3.1.2 Evolving process

The cooling crystallization of an Active Pharmaceutical Ingredient was monitored through *in situ* Infrared spectroscopy. This case study was proposed to develop PCA-based approaches for transient data. Although PCA is used to track evolving systems and for off-line kinetic modeling (e.g., Alcalà et *al.*, 2010), it encounters limitations during on-line monitoring of these systems (De Ketelaere et *al.*, 2015), since it detects excessive false alarm (e.g. observations misclassified as out-of-control). Moving Window Principal Component Analysis (MWPCA) is frequently used for on-line monitoring of evolving processes, since it adapts PCA model such that the process dynamics can be taken into account (Jeng, 2010). It is able to reduce the number of false alarms with the respect of the conventional PCA.
Nevertheless, MWPCA seemed rarely applied to spectroscopic data for the on-line monitoring of evolving systems. In this work, T^2 and Q and contribution plot based on MWPCA were employed to detect the nucleation and identify the spectral variables that were changing due to nucleation.

3.2 Approaches to investigate phenomena occurring in evolving systems

3.2.1 Treatment of hyperspectral data

The system investigated was a surfactants paste that dissolves with water. Since data coming from confocal Raman microscopy vary over the time and along the space, their interpretation is not always straightforward. Multivariate Curve Resolution (MCR) is a very popular method for the resolution of spectra and for the determination of the evolution of the components over the time and/or along the space (Tauler, 1995; de Juan et *al.*, 2014). It leads to more physically meaningful and interpretable results than PCA and EFA. To this aim, MCR was used to infer information about the dissolution behavior of the surfactants present in the paste.

3.2.2 Treatment of X-ray Powder Diffraction data

This case study was focused on the understanding of the setting reaction of K-struvite, an innovative cementing material. The goal was the estimation of the conversion curves of the different species and the extraction of their diffraction patterns from in situ X-ray Powder Diffraction data.
Although Multivariate Curve Resolution (MCR) is a useful tool, the results depend on the quality of the initial estimates since it is an iterative approach. On the other hand, STOCSY (Cloarec et *al.*, 2005) can represent a efficient alternative to estimate diffraction patterns, since spectra are characterized by fixed proportionality between peaks belonging to the same species, as in Nuclear Magnetic Resonance spectra. However, it may find difficulties in estimation when dealing with evolving system, where reactant peaks can overlap with the products ones. Particularly, Time Window STOCSY (TWSTOCSY) was proposed here as a method to estimate crystalline patterns for evolving systems when peaks overlap. Thus, a combined procedure was suggested to estimate patterns and the evolution of phases during the reaction of K-struvite: TWSTOCSY that could provide more accurate initial guesses for MCR implementation to improve the estimation of spectra and evolution of species. This procedure required a limited a-priori knowledge of the spectra and species involved in the reaction.

System investigated	Aim	Spectroscopic technique	Multivariate Technique
Commercial detergents	Quality control	Infrared	• PCA • PLS-R
Active Pharmaceutical Ingredient	Crystallization monitoring	In situ Infrared	• PCA • Moving Window PCA
Cementing material	Reaction monitoring	In situ X-ray Powder Diffraction	• EFA • MCR • Time Window STOCSY
Surfactants paste	Dissolution investigation	Confocal Raman Microscopy	MCR

Table 2 - Overview of the case studies investigated, the spectroscopic techniques and the multivariate techniques employed in this thesis.

Chapter 4
Quality monitoring of commercial detergents

Liquid commercial hard surfaces detergents are a complex blend composed by different chemical species, whose quality strongly depends on the relative proportions. Quality control is usually performed on the end-product at the completion of the process with off-line conventional analytical techniques (e.g. chromatography) that introduce time delay and consequently reduce the effectiveness of quality control. Indeed, in case the product does not meet certain specifications, it has to be reprocessed, implying an increase of cost and time waste. Thus, for the analysis and control of critical quality variables (i.e. the compounds proportions) during the manufacturing process, real time analyzers are required. For the case at hand, a proper experimental tool might be the attenuated total reflectance (ATR) coupled with Fourier transform infrared (FTIR) spectrometer (Stuart, 2004). This is an innovative, non-destructive analytical technique, capable of measuring in very fast times aqueous samples, characterizing materials in a really efficient way and that are well suited for on-line measurements. Nevertheless, because of the large amount of spectral information, interpretation and correlation of the collected spectra with quality variables is a challenging task.
In this thesis, we tackle the issue of the on-line monitoring of commercial detergent mass production through a method based on a multivariate statistical process control approach applied to FTIR measurements of liquid detergent samples. Two different scenarios will be considered for this purpose: on one hand, the detection of samples that are characterized by a higher concentration value for one component of the mixture. On the other hand, the determination of the concentration of some selected compounds and an approach is proposed to detect *out-of-control* samples, particularly to assess the robustness of the estimation when deviations of other compounds are not taken into account during the model calibration.

4.1 Experimental

4.1.1 Materials

Different batches of liquid commercial detergent were generated mixing the following components: sodium hydroxide, pH buffer, chelating agents, anphoteric surfactant, ethanol, fatty acid, non ionic surfactant, sodium

carbonate, perfume, polymer additive. Two datasets were produced where the compounds concentration in each sample is jointly varied according to an *I*-optimal design (Montgomery, 2008a).

4.1.2 Infrared measurements

The infrared measurements were performed at the Procter & Gamble R&D Brussels Innovation Center with a Thermo Scientific Nicolet™iS™10 FT-IR Spectrometer with a deuterated triglycine sulfate (DTGS) detector and a KBr/Ge mid-infrared optimized beam-splitter. The spectra cover the range from 3000 to 800 cm^{-1} with a wavenumber resolution equal to 1.928 cm^{-1}.

4.1.3 Dataset for on-line detection models

This dataset is generated to detect out-of-control samples that can be characterized by the concentration of one component higher than the nominal value during detergent production.
Samples that respect the standard of the end-product were generated and hereafter they will be referred as normal operating conditions (NOC). The in-control set consists of 71 samples were randomly chosen and are characterized by fluctuations from their nominal values as summarized in Table 3. The training data set includes 53 spectra (75 % of the in-control dataset, randomly chosen). The remaining 18 spectra were subsequently used as test set for the model validation. The corresponding spectra are depicted in Figure 3a and b. Other 44 samples (spectra in Figure 3c) were generated with the same average values and fluctuations, but with a concentration of the anionic surfactant 15 % greater than the nominal value assumed in the training set. Therefore, these latter samples may be regarded as an out-of-control dataset when compared with the former one. In all the samples, small variations in the components are introduced in order to mimic typical fluctuations unavoidably present in the standard mass production. The deviations from the nominal values are summarized in Table 3.

Compounds	Deviation from the nominal value	
	in-control samples	**out-of-control samples**
anionic surfactant	± 5 %	± 15 %
▪ **sodium hydroxide** ▪ **pH buffer** ▪ **chelating agents** ▪ **anphoteric surfactant** ▪ **ethanol** ▪ **fatty acid** ▪ **non ionic surfactant**	± 10 %	± 10 %
▪ **sodium carbonate** ▪ **perfume**	± 14 %	± 14 %
polymer additive	± 25 %	± 25 %

Table 3 – Deviations from the nominal values for the compounds present in the in-control and out-of-control samples set.

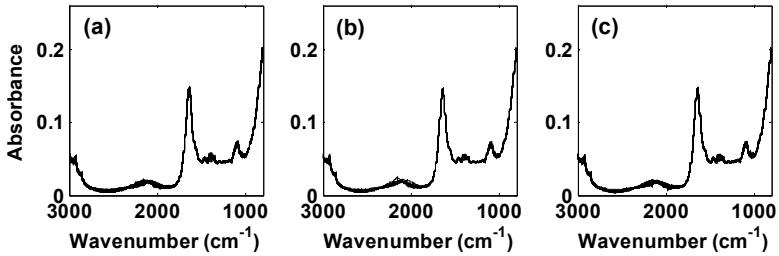

Figure 3 - FTIR spectra for the: (a) training data (53 samples); (b) test data (18 samples); (c) out-of control data (44 samples).

4.1.4 Dataset for on-line estimation of compounds concentration

This dataset is generated for the on-line monitoring of detergent mass production as the previous one. However, here the goal is the determination of the concentration of some selected compounds (sodium hydroxide and non-ionic surfactant) and development of methods to detect *out-of-control* samples where the concentration of anionic surfactant

deviates from the nominal one and is not taken into account in the model calibration.

In-control samples were generated where sodium hydroxide and non-ionic surfactant concentration jointly vary (± 10 % of the average nominal values), while anionic surfactant concentration was kept at low level. Particularly, 34 samples were randomly selected from the in-control set to generate the training set to calibrate the model, the 6 left samples were considered for the validation set. The IR spectra are depicted in Figure 4a and Figure 4b, respectively. In addition, a further set was generated which is an out-of-control test set (12 samples), characterized by anionic surfactant concentration 22 % higher than the NOC value (Figure 4c), while variations of sodium hydroxide and non-ionic surfactant concentration for these samples were the same designed for the in-control samples. Concentrations of other compounds (sodium carbonate, fatty acid, pH buffer, chelating agents, anphoteric surfactant, ethanol, perfume, polymer additive) were slightly varied in all the samples in order to simulate industrial process fluctuations. The deviations from the nominal values are summarized in Table 4.

Compounds	**Deviation from the nominal value**	
	in-control samples	**out of-control samples**
▪ **sodium hydroxide** ▪ **non ionic surfactant**	± 10 %	± 10 %
anionic surfactant	low level	+ 22 %
sodium carbonate	± 14.3 %	± 14.3 %
perfume	± 14 %	± 14 %
polymer additive	± 13.6 %	± 13.6 %
▪ **pH buffer** ▪ **chelating agents** ▪ **anphoteric surfactant** ▪ **ethanol** ▪ **fatty acid**	± 10 %	± 10 %

Table 4 – Deviations from the nominal value for in-control and out-of-control samples.

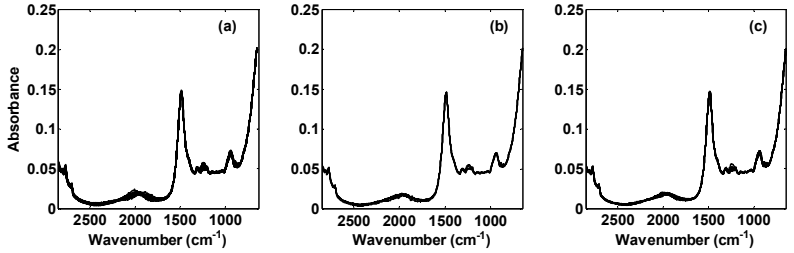

Figure 4 - Infrared spectra for (a) training (34 samples), (b) validation (6 samples) and (c) out-of-control (12 samples).

4.2 Methods

4.2.1 Elliptical NOR for detection of out-of-control samples

Traditionally, on-line process monitoring requires the analysis of T^2 and Q on separate scalar charts (see e.g. Borin & Poppi, 2007 for an application to FTIR measurements). Eventually, one defines the normal operating region (NOR) of the regular process conditions in a bivariate plot (Romagnoli & Palazoglou, 2012) as the square region: $[T^2 - Q] \in [0 - T^2_{\text{lim}}] \times [0 - Q_{\text{lim}}]$. Recently, different techniques have been proposed to combine the two metrics into a unified statistic (Qin, 2003), and through kernel density estimation (e.g., Chen et *al.*, 2004) that allow to reduce the work load of plant operators. Chen et *al.* (2004) proposed a joint estimation of these two statistics by resorting to a nonparametric evaluation of the bivariate probability density function estimated from the observed values for the scalars T^2 and Q.

From a theoretical point of view the T^2 follows a generalized student distribution and the Q statistic a Chi-Squared distribution (Jackson, 1991). As a result, the in-control region defined by T^2 and Q is the joint of two ellipsoids (Qin, 2003). Hence, the aim of this work is the estimation of a new joint probability density function that is closer to these ellipsoids as much as possible. Thus, an alternative, simple approach is here proposed to (i) evaluate a joint Gaussian probability density function of these two new scalars and (ii) estimate a new bounded region.

Since it requires the variables to follow a Gaussian dispersion, the approach is based on a nonlinear transformation $\mathbf{z}^*=[T^{2*}, Q^*]$ of the original scalars. The final goal is to assess whether it is possible or not to correctly detect the anomalies and properly classify the samples.
Thus, in case these statistics do not follow the distributions assumed in the theory, we address a Box-Cox transformation for the statistics:

$$z_i^* = \begin{cases} \left(z_i^{\gamma} - 1\right)/\gamma & \gamma \neq 0 \\ \log z_i & \gamma = 0 \end{cases} \qquad i = 1,2, \quad z_i = T^2, Q \tag{38}$$

The goal is to approximate the data as much as possible to a Gaussian variable. The value of γ in Equation (38) was found by maximizing the Akaike Information Criterion (Akaike, 1974). The normality assumption for the transformed data is finally tested performing a Lilliefors goodness-of-fit test (Lilliefors, 1967). Thus, the bivariate samples of the statistic $\mathbf{z}^*$ can be regarded as outcomes of the multidimensional Gaussian random variable:

$$f_Z\left(\mathbf{z}^*\right) = \frac{1}{\left(\sqrt{2\pi}\right)^2 \sqrt{\det \mathbf{V}^*}} \exp\left(-\frac{1}{2}(\mathbf{z}^* - \overline{\mathbf{z}}^*)^T \cdot (\mathbf{V}^*)^{-1} \cdot (\mathbf{z}^* - \overline{\mathbf{z}}^*)\right) \tag{39}$$

where the term:

$$g\left(\mathbf{z}^*\right) = (\mathbf{z}^* - \overline{\mathbf{z}}^*)^T \cdot (\mathbf{V}^*)^{-1} \cdot (\mathbf{z}^* - \overline{\mathbf{z}}^*) = const \tag{40}$$

represents isolevel curves identified in the $\mathbf{z}^*$ space. The $\mathbf{V}^*$ matrix is the covariance matrix estimated from the two transformed statistics evaluated for the samples. Equation (40) allows to define a recipe for the calculation of a new normal operating region that can be adopted for the statistical control: in detail, we will refer to the ellipse in the $[T^{2*}\text{-}Q^*]$ space including the most of the observed $\mathbf{z}^*$ values (e.g. the 95% of the data). This region will be referred as the Elliptical Normal Operating Region (ENOR).

4.2.2 PLS-R for on-line estimation of compounds concentration

Partial Least Squares Regression (PLS-R) is considered as the best linear multivariate technique for the quantitative analysis of spectroscopic data because it enables to overcome common problems such as collinearity, band overlaps and interactions.
Different works in literature usually determine the best calibration PLS-R model for compounds concentration in detergents formulations (Rohman et *al.*, 2011), however the possible presence of external interferences is not considered. As a consequence, prediction ability could be worsened. In fact, since the PLS-R model is built only on a limited number of compounds, it could be no longer consistent when the system is out-of-control, that is in presence of large deviations of other compounds not taken into account in the PLS-R model calibration. Hence, the PLS-R model could lead to wrong conclusions when this scenario occurs.
Here, a Q_x statistic is proposed to assess the concentration prediction reliability when a fault occurs, i.e. a perturbation from the NOC.

4.2.2.1 PLS-R based statistical process control

Beside the quantitative estimation of the compounds (see section 2.6), a PLS-based monitoring technique was here applied (Kourti, 2005). An important factor to consider is the consistency of the prediction of unknown samples, indeed the new observations should lie in the calibration region for a correct concentration prediction. Different works are devoted to this topic (Bu et *al.*, 2013; Pierna et *al.*, 2002). Here, the on-line detection of deviations from nominal conditions and, as a consequence, the accuracy of the concentration prediction was performed by resorting to the Q_x statistic (Li et *al.*, 2010). Such statistic (Godoy et *al.*, 2014) is calculated for the *k*-th sample spectrum according to the Equation (41).

$$Q_{x,k} = \|\boldsymbol{x_k} - \widehat{\boldsymbol{x_k}}\|^2 = \left\|\boldsymbol{x_k} \cdot \left(\mathbf{I} - \mathbf{P_A} \cdot \mathbf{R^T}\right)\right\|^2 \quad (41)$$

Where the subscript x in the Q refers to the residual of the data matrix **X**, $\widehat{\boldsymbol{x_k}}$ is the *k*-th spectrum as predicted by the PLS model considering the

first A latent variables. The threshold limit for this statistic can be evaluated as reported in Equation (15) or (16). Then, the procedure can be summarized as represented in Figure 5: the spectra are collected and PLS model is implemented, if the Q_x exceeds the threshold estimated for the in-control sample set, the model is not suitable anymore for the prediction of compounds concentration.

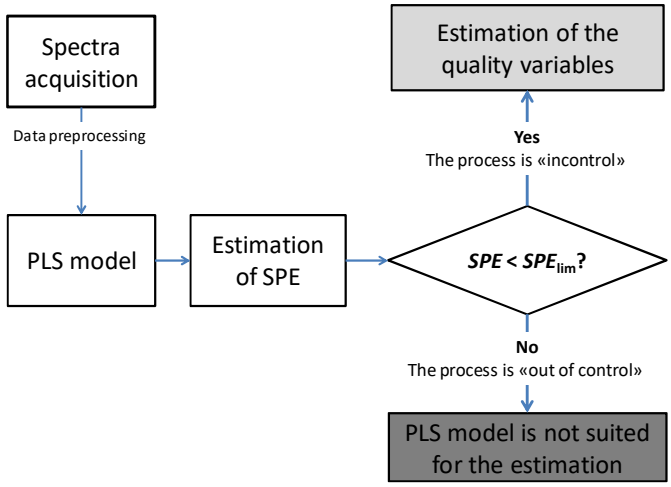

Figure 5 – Procedure for on-line monitoring of compounds concentration based on PLS-R and Q_x statistic.

4.3 Results

4.3.1 On-line detection of out-of-control samples

The PCA model was identified using the training data set based on the Equations illustrated in section 2.4. According to the cumulative variance criterion, the appropriate number of principal components in the PCA model was found to be A=6, explaining 91.1% of the total data variance, while the other components explain less than the 1 % and can be thus discarded (Figure 6).

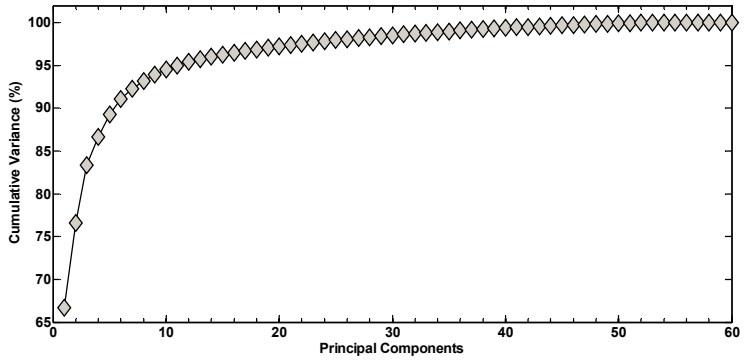

Figure 6 – Cumulative variance explained of the matrix **X**.

The samples of the T^2 and Q statistics for the training, in-validation and out-validation set were then generated by applying Equations (12) and (14) described in section 2.5. The traditional bivariate chart T^2 and Q statistics for process control is depicted in Figure 7. As it can be noted, the samples belonging to the training set (open circles) appear as the joint of two ellipsoids.

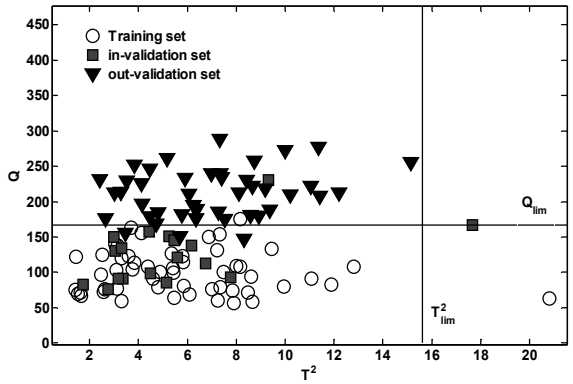

Figure 7 - Scatter plot of the T^2 and Q statistics evaluated for the observations belonging to the training data set, test set and out-of-control data set. Dashed-dotted lines are the limiting values calculated for the statistics.

From a theoretical point of view the T^2 and Q statistics are independent of each other (Chen et *al.*, 2004; Jackson, 1991). Therefore, the independency of the two statistics is verified through the covariance matrix of the statistics estimated for the training set shown below and the correlation coefficient that is equal to -0.109.

$$V = \begin{bmatrix} 12.16 & -11.48 \\ -11.48 & 911.64 \end{bmatrix}$$

In order to consider a joint distributed function, the variables should follow a normal distribution. Since T^2 and Q statistics generally follow a generalized student distribution and a Chi-Squared distribution, respectively, before the estimation of the joint probability density function, the normality of these statistics is tested. To this aim, the Lilliefors test (Lilliefors, 1967) is carried out. In fact, it is found that normality assumption for Q was rejected with a p-value=0.021. While for T^2 is equal to 0.192, therefore it can be considered as following a Gaussian distribution. The bivariate sample of statistics $z^*=[T^2, Q^*]$ can be reasonably approximated as an outcome of a multidimensional Gaussian random variable through the Box-Cox transformation as proposed in section 4.2.1. From the maximization of the Akaike Information Criterion, it was found that $\gamma=0$ for the SPE statistic (i.e. $Q^*=\log Q$) as depicted in Figure 8 where the Lilliefors tests give a p-value equal to 0.123.

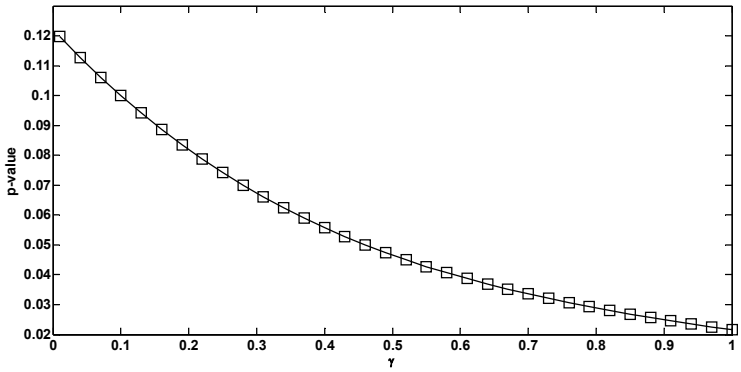

Figure 8 – P-value estimated from the Lilliefors test for the Q statistic.

A summary of the results is reported in Figure 9, that reports the scatter points of the T^2 and Q^* statistics as they were estimated for the different data set: white circles, grey squares and black triangles correspond to the training set, test set, and out-of-control set, respectively. For the sake of completeness, it is shown in the Figure: (i) the ENOR calculated with Equation (40) and including the 95 % of the training data set (dotted line) and (ii) the threshold values (dashed-dotted line) estimated for the statistics T^2_{lim}=15.61 and Q^*_{lim}=5.11 as they were evaluated through the

expressions available in the literature. Some comments are in order. Both the traditional NOR and the ENOR correctly classify almost all the samples of both the training (NOR: 52 out of 53 samples, that corresponds to 96%; ENOR: 51 out of 53 samples that corresponds to 92%) and the test set (both NOR and ENOR: 16 out of 18 that is 89%) as true negative (i.e. $0<Q^*<Q^*_{lim}$ and $0<T^2<T^2_{lim}$). It should be remarked that the successful classification of the test set further confirms the properness of the PCA model with A=6 principal components (at least for the variations of the compound here explored). Furthermore, the additional samples obtained at the higher level of anionic surfactant concentration fall almost completely out of both the operating regions and they are correctly classified as abnormal process conditions. An effective separation between the in-control (both training and test set) and out-of-control data is again observed. As a final remark, one can appreciate from visual inspection that the Q^* statistic is also able to successfully discriminate the two data sets, at least for the case at hand. Whereas, no deviation from the nominal behaviour can be noticed through T^2. Indeed, as also reported by Qin (2003), the T^2 is less sensitive to deviations than Q statistic. Although the anionic surfactant is also present in the in-control set, the deviation of its concentration from the nominal value (+ 15%) in the out-of-control set breaks the correlation between the spectral variables existing for the training set. This means that the projection of all the spectral variables onto the PCA model are still near to the origin of the PCA subspace, thus T^2 values are less than the limit. However, since only a subset of spectral variables (in this case related to the anionic surfactant) is deviating, the Euclidean distance from the projections becomes high and Q statistic exceeds the threshold. This occurrence can be compared to the red sample depicted in Figure 1 (section 2.5.1).

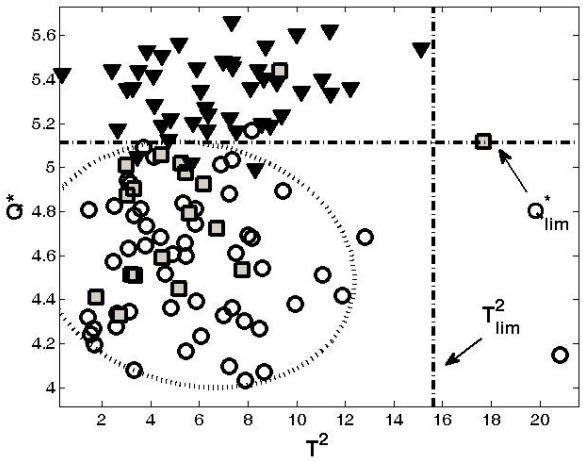

Figure 9 - Scatter plot of the statistics T^2 and Q*. White circles represent the observations for the training data set, gray squares the observations for the test set, black triangles the out-of-control data set. Dashed-dotted lines are the limiting values calculated for the statistics. The dotted ellipse is the ENOR evaluated with Equation (40).

The efficiency of the protocol here introduced is confirmed by representing the Receiver Operating Characteristic (ROC) curves (Scheipers et al, 2005), that are two-dimensional graphs of the true positive rates (TPs; i.e., successes) versus the false positive rates (FPs; i.e., false alarms). To perform the ranking statistical test, a scalar metric is required. In this work, we consider two scalars for the three datasets: (i) the usual Q statistic and (ii) the distance d_e for the ENOR defined in Equation (42).

$$d_e = (\mathbf{z}^* - \overline{\mathbf{z}}^*)^T \cdot (\mathbf{V}^*)^{-1} \cdot (\mathbf{z}^* - \overline{\mathbf{z}}^*) \tag{42}$$

Where the matrix **V*** is the estimated covariance matrix of the samples **z***. The results are reported in Figure 10. The area under the ROC curve is the so-called AUC index, which is a scalar measure of the overall performance of a classifier, averaging across different thresholds that can be used to generate a classifier. In general, a model with a larger AUC is preferred to a model with a smaller one. The AUC of a random classifier is 0.5, whereas AUC=1 corresponds to perfect classification.

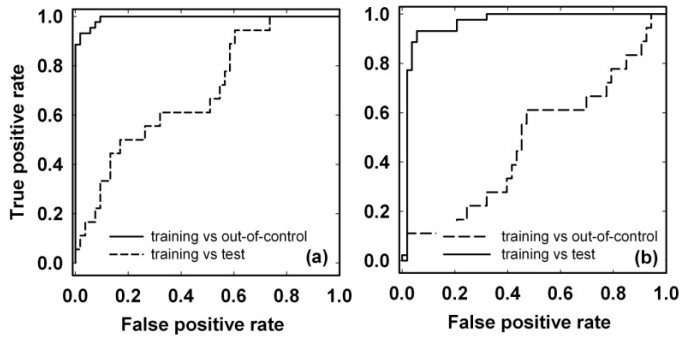

Figure 10 - ROC curves resulting from the comparison of the training with out-of-control data (solid line) and the training with test data (dashed line) for (a) the Q statistic and (b) the d_e scalar.

The AUC values determined are shown in Table 5, where the comparison among the three different data sets, together with the related coefficient intervals estimated with a bootstrap technique are reported. The ideal scenario would be: (i) an AUC value as close as possible to 1, when comparing the training set with out-of-control set (thus meaning a perfect separation between the two classes) and (ii) an AUC value close to 0.5 when comparing the training set with the test set. It is further confirmed how the proposed procedure shows a high capability to distinguish the in-control from the out-of-control data.

		AUC	AUC_{min}	AUC_{max}
Q statistic	Out of control	0.994	0.974	0.999
	Test set	0.696	0.553	0.826
d_e scalar	Out of control	0.962	0.881	0.988
	Test set	0.482	0.339	0.656

Table 5 - AUC scalars for the Q statistic and the d_e scalar.

In conclusion, it was found, at least for the case under investigation, that the proposed protocol correctly classifies the samples with a performance, at least comparable, with the traditional bivariate plot of T^2 and Q statistics (Figure 7), but with a slightly higher specificity, since the test set is classified as belonging to the training set (AUC=0.482<0.696). As a final remark, it should be noticed that, up to our knowledge, a PCA based statistical control has been seldom implemented in the framework of infrared spectroscopy measurements for detergent quality monitoring.

4.3.2 On-line estimation of compounds concentration

The PLS-R model (see Equations reported in section 2.6) was built on the training set samples represented by the experimental matrix $\mathbf{X}_{(34\times1142)}$ and concentration matrix $\mathbf{Y}_{(34\times2)}$. Model matrices (**P**, **Q** and **B**) are evaluated using SIMPLS algorithm (De Jong, 1993). For the case at hand, six latent variables were chosen as the variance explained for both **X** and **Y** achieves 92 and 97 %, respectively. The subsequent components explain less than the 1 % and can be thus discarded (Figure 11).

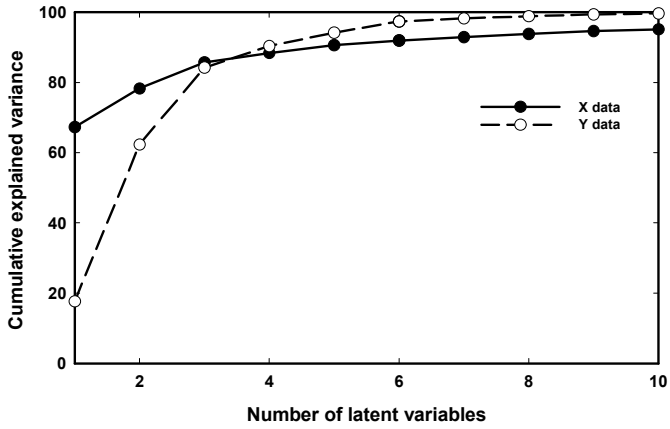

Figure 11 – Cumulative explained variance for the data matrix **X** and **Y**.

According to the procedure proposed in section 4.2.2.1, the Q_x statistic for each sample and the $Q_{x,lim}$ were calculated according to Equation (41) and (16) and they are reported in Figure 12. It can be observed that samples belonging to training and validation sets have Q_x values smaller than the threshold. Thus, they are correctly classified as in-control, that is the PLS-R model is supposed to correctly predict the quality variables. On the other hand, the out-of-control samples exceed the limit and anomalous conditions are detected. In these cases, the PLS-R model cannot be used to infer the compound concentrations as suggested in Figure 5.

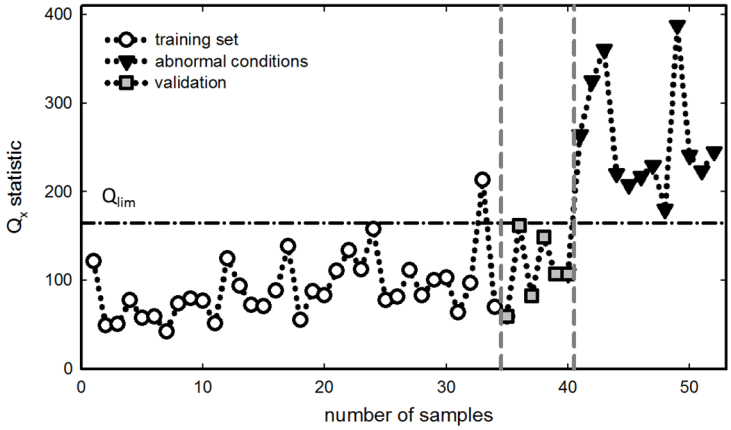

Figure 12 - Q_x statistic control chart. It was evaluated for samples belonging to training (white circles), validation (gray squares) and out-of-control sets (black triangles), respectively. The dashed-dotted line represents the Q_x limit.

In order to show the prediction ability of the PLS-R model for the three different samples sets (training, validation and out-of-control test sets), the sodium hydroxide and non-ionic surfactant concentration ($\mathbf{y}_k$) were calculated according to Equation (22).

The PLS2 model here developed demonstrates high predictive performance achieving R^2 values of 97.4 and 97.3 % for sodium hydroxide and non ionic surfactant concentration estimation (for the training set), respectively. Similarly, the Root Mean Squared Error of Calibration (RMSEC) values are quite low and equal to 0.077 and 0.079. For validation set the Root Mean Squared Error of Prediction (RMSEP) is equal to 0.28 and 0.16. The results are summarized in Figure 13 where the experimental vs predicted concentrations (arbitrary units) for the three different sets are represented. In more detail, Figure 13a and Figure 13b refer to the sodium hydroxide concentration and non ionic surfactant concentration, respectively. It can be seen that the training samples (white circles) and the validation data (gray squares) are well predicted for both sodium hydroxide and non ionic surfactant. On the other hand, it was observed that the out-of-control samples, when projected onto the PLS model (black triangles in Figure 13), cannot be accurately predicted. In particular, the sodium hydroxide concentration was underestimated, whereas the non-ionic surfactant was slightly overestimated. The explanation of such lack of fit seems obvious: possible variations of anionic surfactant concentration were not included into the PLS-R model

calibration. As a consequence, the model is not suited to predict concentrations corresponding to out-of-control samples.

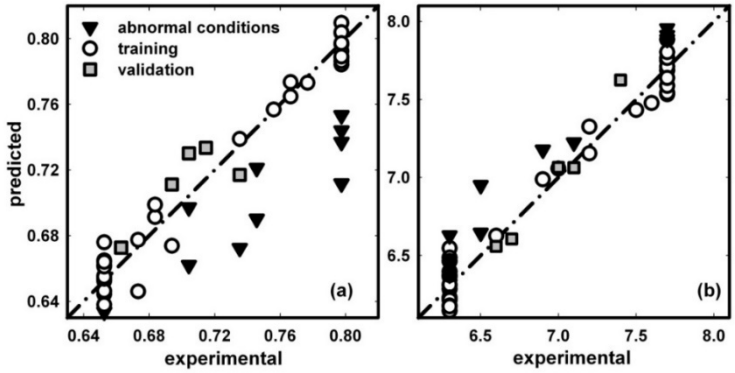

Figure 13 - Experimental vs predicted concentration of sodium hydroxide (a) and non-ionic surfactant (b). White circles, gray squares and black triangles represent training, validation and out-of-control samples, respectively.

The efficiency of the Q_x statistic is further evaluated by means of the Receiver Operating Characteristic (ROC) curves (Scheipers et *al.*, 2005). Here, two ROC curves were determined as in the previous case study: (i) training set was compared with out-of-control set and (ii) training with validation set and depicted in Figure 14. The obtained AUC values for cases (i) and (ii) were AUC_1=0.989 and AUC_2=0.77. This confirms the capability of the Q_x statistic to distinguish the in-control from the out-of-control samples.

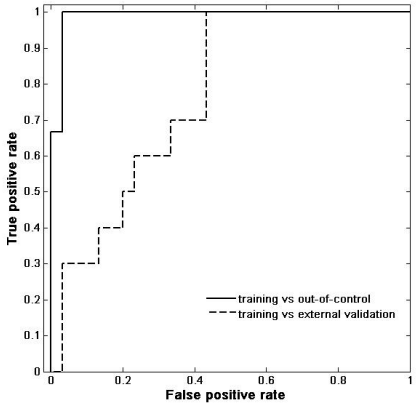

Figure 14- ROC curves resulting from the comparison of the training with out-of-control data (solid line) and the training with validation data (dashed line) for the Q_x statistic.

Results demonstrated that PLS model correctly estimates the compounds concentration when the system is in-control. In addition, the Q_x statistic was demonstrated as an effective tool to clearly detect the out-of-control samples. This can be explained considering that the model was calibrated based on a specific correlation between the spectral variables. As a consequence, the deviation of anionic surfactant concentration from the nominal value (+ 22%) in the out-of-control samples is not consistent anymore with that correlation, since only a subset of spectral variables (in this case related to the anionic surfactant) is deviating. This means that the spectral variables belonging to the out-of-control samples moved off the plane defined by the latent variable and the statistic become higher than the limit. This occurrence is similar to the case of the red sample depicted in Figure 1 (section 2.5.1). Therefore, Q_x statistic can give information about the consistency of the PLS model. Indeed, it cannot be employed to estimate the quality variables in case of operating conditions far from the NOC.

Chapter 5
Monitoring of cooling crystallization of Isonicotinamide

This case study is focused on the monitoring of cooling crystallization of Isonicotinamide (INA) in various solvents. It is a pyridine derivative with an amido group in γ-position (INA molecular structure is shown in Figure 15) and has anti-tubercular, anti-pyretic and anti-bacterial properties.

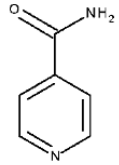

Figure 15 – Molecular structure of Isonicotinamide.

This Active Pharmaceutical Ingredient (API) is a popular coformer that can be used as partner molecules with Active Pharmaceutical Ingredients (APIs) in co-crystal preparation (Filho et *al.*, 2006). Moreover, several metal complexes of Isonicotinamide have been used as drugs in medicinal processes since some coordination compounds of this relevant biological ligand are more effective than the free Isonicotinamide molecule (Yurdakul et *al.*, 2003). Since its use in drug industry are becoming relevant, proper tools are required to in-line monitor crystallization of INA.

Its industrial production is usually carried out by crystallization. Product properties, i.e., crystal shape, crystal size distribution (CSD), mean crystal size and the polymorphic form may be affected by different factors (temperature, concentration of the solute, presence of seeds or impurities). For these reasons, monitoring and control batch cooling crystallization processes plays a key role in optimizing product quality and process performances. To this aim, the U.S. Food and Drug Administration (FDA) initiative promotes the use of in situ analytical technologies, usually referred to as Process Analytical Technologies (PATs) with advanced control methodologies for process understanding, analysis, and control (U.S. Food and Drug Administration, 2004). While, other industries had already employed these tools, pharmacy industry had found difficulties in their introduction for process monitoring, due to rigid regulatory systems.

During cooling crystallization, the API is completely dissolved in the solvent at temperature higher than the saturation temperature, where the solubility is relatively high. When the solution is cooled down, the system reaches supersaturated conditions, where the solute concentration exceeds the solubility at that temperature. Then, the nucleation of new crystals

begins. However, the temperature at which it occurs depends on different factors such as initial concentration, cooling rate, level of agitation and impurities (Nývlt, 1985). Hence, since nucleation is a significant stage that can affect these properties during crystallization, the detection of the exact moment corresponding to the formation of the very first crystals in the system is essential.

Over the last decades, Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy has revealed useful for in-line process monitoring, control of solute concentration, supersaturation and nucleation detection in batch cooling crystallization (Schaefer et *al.*, 2013, Thirunahari et *al.*, 2011; Chen et *al.*, 2009). Nevertheless, since these real time analyzer provide highly informative data, methodologies able to interpret and gather information from these data are required.

The aim of this case study is to correctly detected nucleation as spectra are collected and identify which spectral variables are mostly changing. Therefore, firstly the off-line detection of the nucleation temperature pursued through multivariate technique is assessed, then a procedure is developed and tested to take into account the transient nature of the data.

5.1 Experimental

5.1.1 Materials

Due to issues with availability, INA was purchased from two vendors, Acros Organics and Sigma-Aldrich. X-ray powder diffraction (XRPD) analysis showed that the batch from Acros Organics was form 2(EHOWIH02) and the batch from Sigma-Aldrich was form 1 (EHOWIH01). Methanol, Acetone, Acetonitrile, Ethyl Acetate, were obtained from Sigma-Aldrich, Dichloromethane was obtained from VWR BDH Prolabo as 99.9% analytical grade solvents.

5.1.2 Experimental setup

The experiments were carried out in the Department of Chemical Engineering, Biotechnology and Environmental Technology (University of Southern Denmark). In-line IR spectra were collected from the INA

solutions during cooling crystallization with a Mettler Toledo ATR FTIR ReactIR 15 equipped with a DiComp Diamond probe every 30 seconds. Each spectrum was recorded covering a spectral range 648.9 - 2998 cm^{-1} with a resolution of 3.73 cm^{-1}.

5.1.3 Cooling crystallization

All cooling crystallizations were carried out using an Easymax 102 synthesis workstation from Mettler Toledo with two 100 mL reactors. Mixing in the reactors was provided with crossbar stirring at 300 rpm and a condenser was mounted on each reactor to recover the evaporated solvent during crystallization. The Easymax 102 synthesis workstation is equipped with a built-in solid state thermostat, which ensured controlled cooling down to -25°C while still maintaining constant cooling rate. Cooling crystallization of INA from each solvent has been conducted with two different initial concentrations, solution saturated at 10°C and at 35°C (30°C for dichloromethane), respectively. Before the cooling started, the solution was heated to 10°C or 15°C above the saturation temperature for 1 hour to ensure complete dissolution of INA. The cooling rate for all experiments was fixed at 0.5°C/min. Once the solution became turbid, the temperature was noted as the observed nucleation point. The cooling was continued until a sufficient amount of crystals nucleated out of solution, or in cases of solvents with low solubility, until the reactor temperature reached -25°C. At this point, experiments were stopped as it was not possible to keep a constant cooling rate below -25°C. IR data were collected using the ReactIR15 probe synchronized with the Easymax temperature control through software and an appropriate peak in the IR spectra was chosen depending on the solvent in order to trend the peak height during experiments.

5.1.4 Dataset for off-line detection of nucleation

This dataset is used for the off-line detection of the nucleation temperature of INA in various solvents. Therefore, spectra right after cooling was initialized until the end of the experiments were analyzed. Since crystallization of INA occurred in various solvents, the models were developed for each solvent separately. For each solvent, the IR spectra

corresponding to solution saturated at low temperature were included in the training set. On the other hand, spectra collected at high saturation temperature and the replicate at low saturation temperature were used as validation.

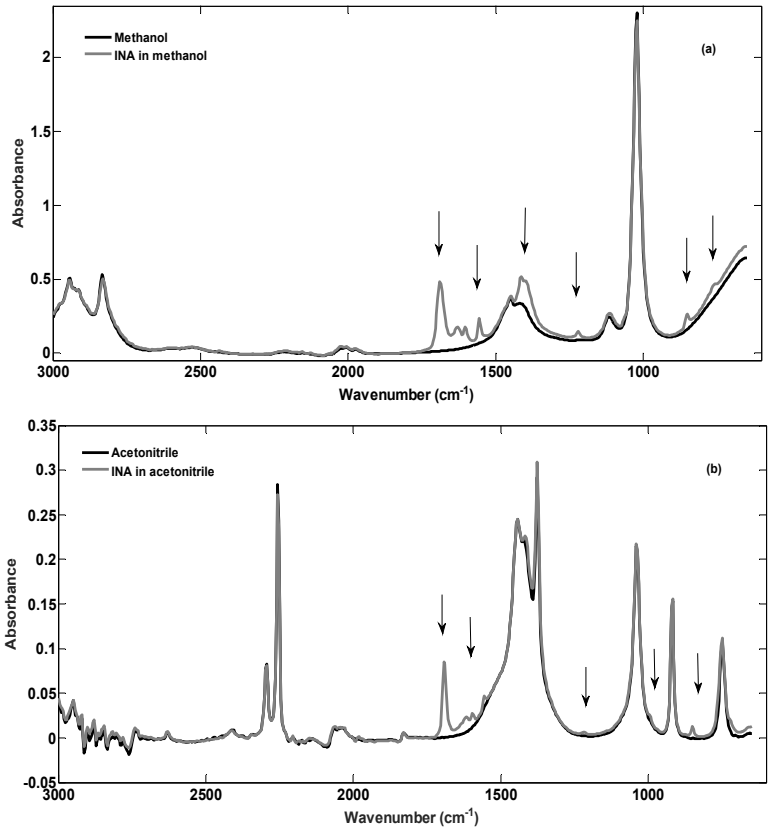

Figure 16 - Infrared spectra of (a) methanol and (b) acetonitrile (black lines) compared with the spectrum of the INA solution collected at room temperature (grey line). INA peaks are indicated with arrows.

Figure 16a depicts the IR spectrum of methanol and INA in methanol. As it can be observed, the main INA peaks are located at 764, 850, 1219, 1413, 1555, 1604, 1630 and 1689 cm^{-1}. While the peaks related to methanol bonds are located at 1022, 1115, 1413 and 1451 cm^{-1}. It should be noted that the INA bond at 1413 cm^{-1} overlaps with the methanol one.
Concerning the IR spectrum of INA solution in acetonitrile, Figure 16b shows that also in this case the main INA peak is located at 1689 cm^{-1}, while the other peaks are distributed in the spectral range 1555 and 1618

cm^{-1}, in addition three peaks are observed at 1219, 992, 850.2 cm^{-1}. The remaining peaks can be assigned to the solvents.

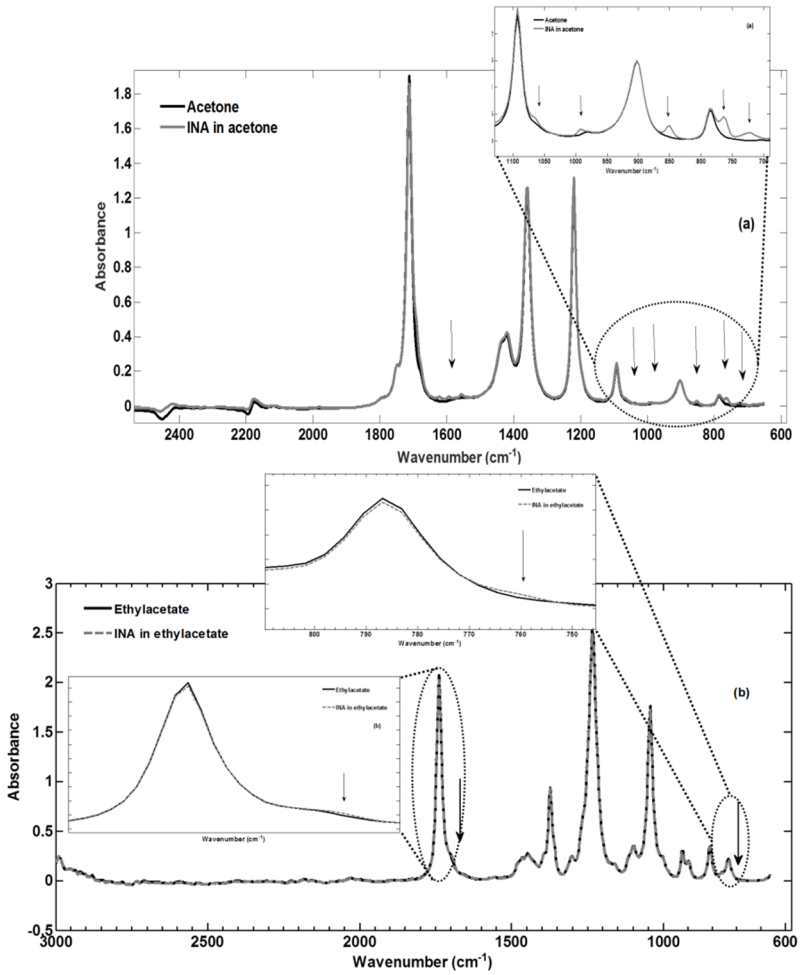

Figure 17 - Infrared spectra of (a) acetone and (b) ethylacetate (black lines) compared with the spectrum of the INA solution collected at room temperature (grey line). INA peaks are indicated with arrows.

While, smaller INA peaks are observed for acetone and ethylacetate in Figure 17a and b, respectively. The peak at 1689 cm^{-1} present for the other solvents cannot be appreciated probably because the solvent peak strongly overlaps with it. However, in case of acetone three peaks are noted quite near (1640÷1525 cm^{-1}), while the other peaks are located at 1067, 992, 850, 764.5, 727.2 cm^{-1}. Regarding the ethylacetate, there is a strong

overlap with the solvent peaks and only two tiny peaks can be distinguished (1700 and 760 cm^{-1}).
As it can be inferred from the previous Figures showing the INA solution in various solvents, that some peaks are located in the same position and they represent specific functional INA group as reported in Table 6.

Spectral region (cm^{-1})	Functional group
1700-1680	Coniugation of C=O with pyridine
1690-1640	C=N stretch
1640-1560	N-H stretch for primary amin
1600 and 1475	C=C aromatic
1350-1000	C-N stretch for amine
900-690	=C-H, out of plane bending
800	N-H, out of plane absorption

Table 6 – Functional groups of INA.

For the sake of brevity, only the spectra collected during the experiment carried out with methanol are reported in Figure 18, spectra pertaining the other solvents show analogous behavior. As expected, the different concentration of INA in the solutions corresponds to higher INA peaks in the spectra (grey lines).

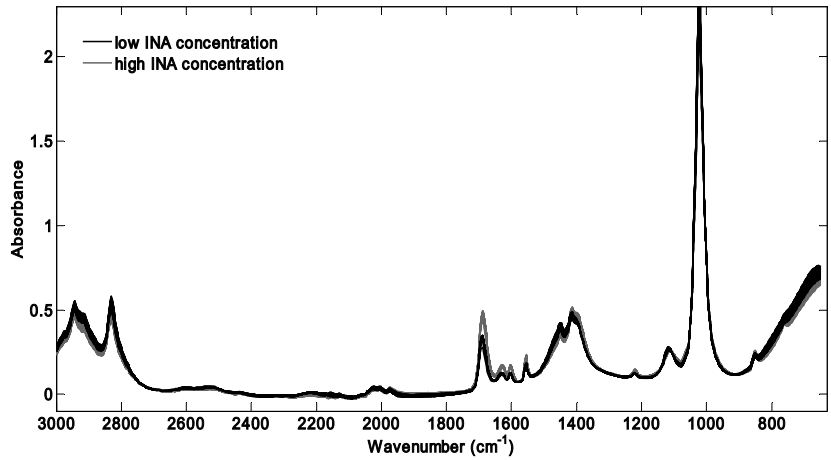

Figure 18 – IR spectra collected during cooling crystallization of INA in methanol. Spectra of the solution at low INA concentration is depicted as black lines, while the low is represented as grey lines.

5.1.5 Dataset for in-line detection of nucleation

This dataset is considered as benchmark to develop and test algorithms for the analysis of transient data. The in-line implementation of these algorithms aims to detect nucleation as spectra are collected and identify which spectral variables are mostly affected by the nucleation. As an illustrative example, only methanol was used. It included spectra collected only during cooling crystallization of INA in methanol, where the solution was saturated at low temperature (10 °C) and depicted in Figure 19.
Therefore, for data processing, we considered only spectra collected from the beginning of cooling (T= 24.91 °C) to the end of experiment (T= -9.78 °C), in the spectral range 648.9÷2002 cm^{-1} in order to exclude contributions not informative for INA crystallization.

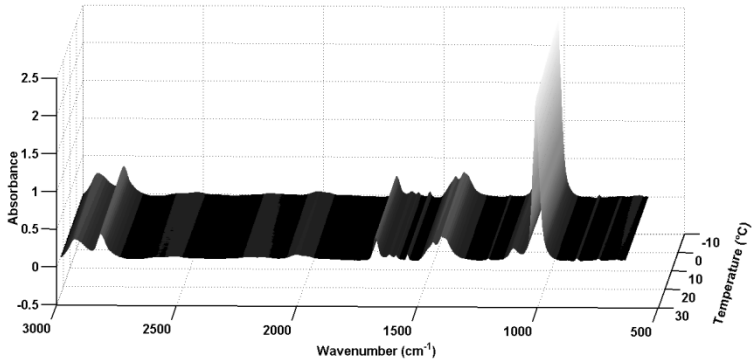

Figure 19 – Infrared spectra collected during cooling crystallization of Isonicotinamide in methanol at low initial concentration.

5.2 Methods

The goal of this case study is the development and assessment of the proper procedures to apply to IR data for the correct detection of nucleation and the identification of spectral variables that are mostly changing during cooling crystallization of Isonicotinamide. Firstly, the off-line detection of the nucleation temperature pursued through static PCA is assessed. Since it is an evolving system, in order to take into account the transient nature of the data, a PCA-based method for the in-line monitoring is developed and tested.

5.2.1 Off-line PCA model

The IR spectra collected during crystallization of Isonicotinamide in various solvents are off-line analyzed through PCA (see section 2.4) in order to give an earlier indication of nucleation occurring and furthermore to help to distinguish the molecular clusters prior to the onset of nucleation.

5.2.2 MWPCA applied to spectroscopic data

In this thesis, MWPCA is proposed to process data coming from spectroscopic measurements during a transient process. Although this procedure is well known and extensively applied in process monitoring (Simoglou et *al.*, 2005; Zhaomin et *al.*, 2014), no works in literature have employed it for the analysis of in situ spectroscopic data along the perturbing variable *t*. Indeed, during a transient process, spectra may reflect step change (behavior different from the starting conditions: a component that disappears, another that appears). They also may be characterized by drifts that do not necessarily imply that the system is out-of-control.

The method here implemented is shown in Figure 20: after selecting the window size, the window was moved along the data, a training set, $\mathbf{X}^{\mathbf{c}}_{(\mathrm{L}\times\mathrm{J})}$, was generated and PCA model was built, that means determining the loading matrix **P**, observations mean and standard deviation, number of principal components, T^2 and Q limits according to Equations (13) and (15). Then, the next two observations were projected onto the PCA model and T^2 and Q were estimated according Equations (12) and (14). The criterion adopted for the moving of the window was developed with the following criteria: if both observations are under the control limits, the new window will be moved forward including these two observations and removing the older two ones. An example of the sliding window is reported in Figure 21, where $\mathbf{X}_c$ and $\mathbf{X}_v$ refers to the training and validation set.

When the statistics evaluated for three consecutive future observations exceed the limits (De Ketelaere et *al.*, 2015), the system is defined as out-of-control, the window does not move forward anymore. In this case, the contribution plot can be analyzed to assess which variables are most

contributing to the statistics. The main advantage of this procedure over the static PCA is that the number of false alarms can be drastically reduced and the detection of the out-of-control state can be more easily identified.

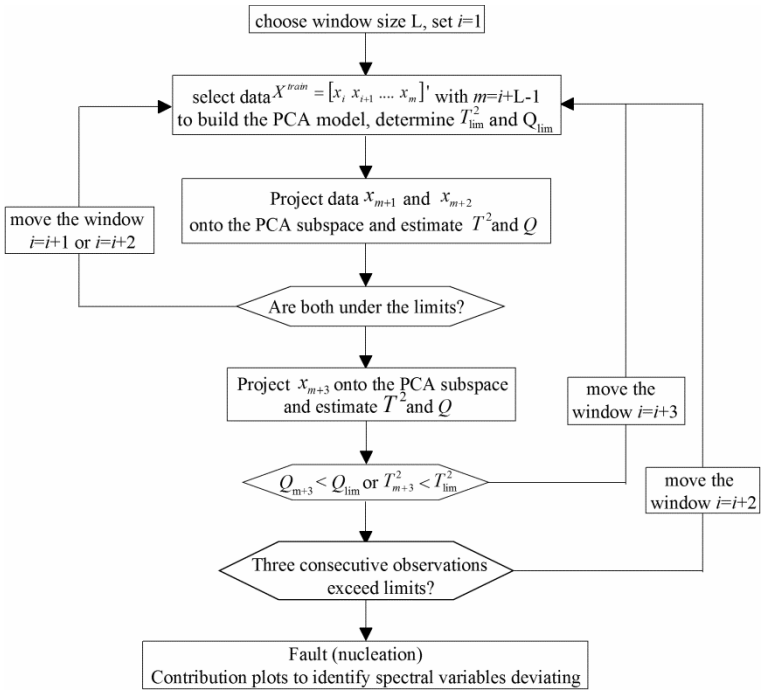

Figure 20 – Moving Window PCA algorithm to monitor evolving system and detect out-of-control observations.

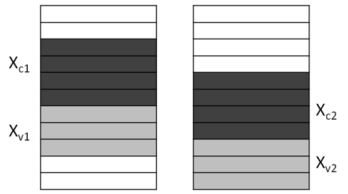

Figure 21 – An example of the sliding window along the data matrix **X** carried out during MWPCA implementation.

5.3 Results

5.3.1 Off-line PCA model

For each solvent, PCA model was built by considering the IR spectra corresponding to solution saturated at low temperature. The PCA model was then validated on the spectra collected at high saturation temperature and the replicate at low saturation temperature. Only the spectra collected right after cooling was initialized until the end of the experiments were taken into account. Spectra were pre-processed through mean-centering and standardization before PCA implementation.
It was found that for the IR data of INA in methanol, the first two scores calculated for the training set (low saturation temperature solution) described the 84.9 % of variation in the data (the first 74.8 %, while the second 10.1 %).

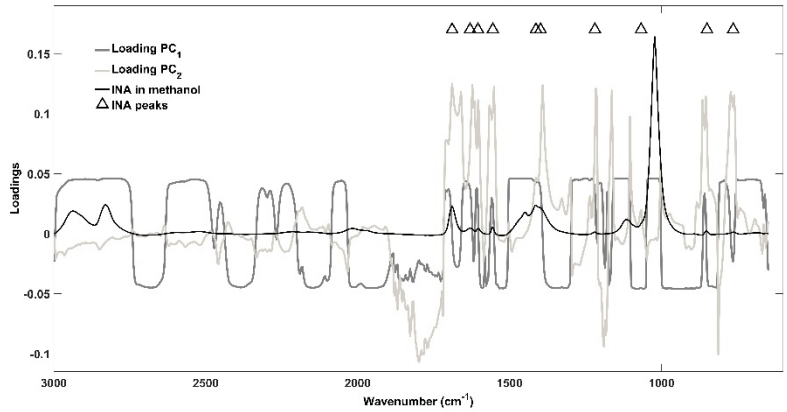

Figure 22 – First and second loading for INA in methanol saturated at 10°C compared with the spectrum of INA in methanol. The main characteristic peaks of the INA are also reported (open triangles).

Figure 22 shows the first two loadings together with the spectrum of INA in methanol (black solid line), where the INA peaks are depicted as open triangles. It appears that the first component (dark grey line) was more influenced by the solvent peaks, the most apparent of which is located at 1020 cm^{-1}. Moreover, the first component also took into account to the temperature variation of the whole spectrum, since a drift of the whole absorbance spectrum towards higher values was appreciated during the

cooling. Regarding the second loading (reported as the light grey line), it was mainly related to the INA peaks (e.g. located at 1413, 1555 and 1689 cm^{-1}) and reflected the significant decrease of INA peaks occurring after the nucleation. Therefore, the crystallization process could be monitored by studying the only second score. Analogous conclusions were inferred from the loading plot obtained for the other solvents, not reported here for the sake of brevity.

Figure 23 reports a phase diagram of the first two scores evolving with respect to the temperature. It appears that the first score significantly increased during the cooling, but for temperatures *T* below - 7 °C it stayed constant. As it regards the evolution of the second score, it slightly increased in the first part of the experiment. Then, it abruptly decreased at temperature T=-7 °C. This qualitative change in the behavior was supposed to be related to the onset of the nucleation process. Thus, analysis using PCA showed that the IR data did give a clear detection of the nucleation points.

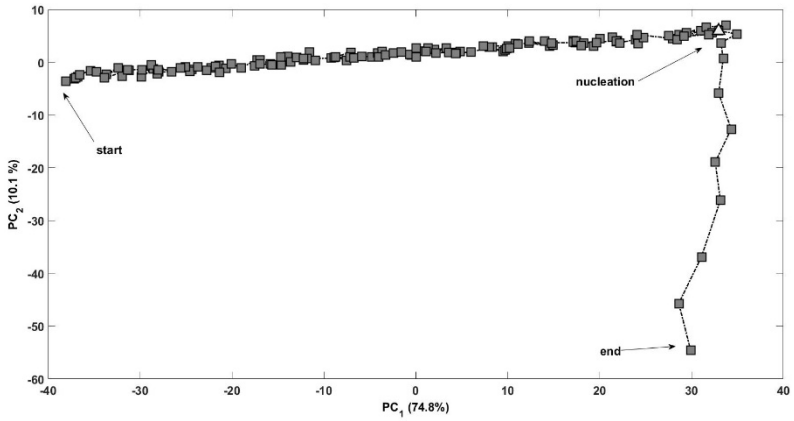

Figure 23 – First and second score for INA in methanol saturated at 10°C.

Concerning the other solvents, the first component described the 38.5%, 56.7%, 51.6%, while the second 14%, 8.7%, 12.9% for acetonitrile, ethyl acetate, and acetone, respectively. In addition, the validation set was projected on the PCA model and the scores were determined. As can be seen from the Figures below, the change of direction of the second score could be associated to the beginning of nucleation for INA in methanol (Figure 24), acetonitrile (Figure 25), ethyl acetate (Figure 26) and acetone (Figure 27). Different results were achieved for the second score of INA in dichloromethane. Due to the low INA solubility in this solvent, the IR

spectrum was characterized by very small INA peaks. As a result, the second score did not show a sharp change of direction as in the case of other solvents (results not reported for sake of brevity). The model and the ability of the score to follow the progress of the crystallization were further confirmed by the validation set represented as dark grey triangles and dark grey diamonds for solution saturated at 10 and 35 °C, respectively. In fact, the temperatures detected through the second score show little variance to the nucleation points observed by the naked eye (represented as light grey squares and diamonds), with variations of ±0.5°C. A little higher difference (±1.3 °C) was obtained in case of INA in ethyl acetate. This could be due to peak overlap and to the fact that the decrease of the INA peaks did not start at the same nucleation temperature visually observed, in particular, the nucleation for the validation set at low and high saturation temperature is detected at 5 and 25 °C instead of 6.3, and 23.7 °C, according to Table 2. Therefore, the PCA could not consistently predict nucleation earlier than that was observed empirically by visual inspection or by raw IR data. However, results matched up with nucleation times that were visually observed. Furthermore, even if the PCA model was built considering spectra at one concentration, the results achieved for a similar system but different concentration were reasonable. In fact, the second score determined for the validation set showed a behavior similar to the score of the training set. This confirms that PCA can be employed to detect nucleation consistently even independent of initial concentrations and, in combination with ATR FT-IR, it revealed as a useful tool for the in-line monitoring of crystallization.

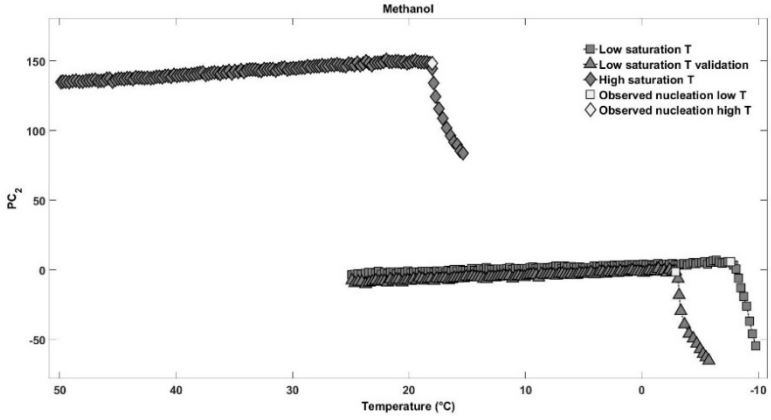

Figure 24 – Second score of INA in methanol compared with visually observed nucleation temperature: training set (10°C) and validation set (10 and 35°C)

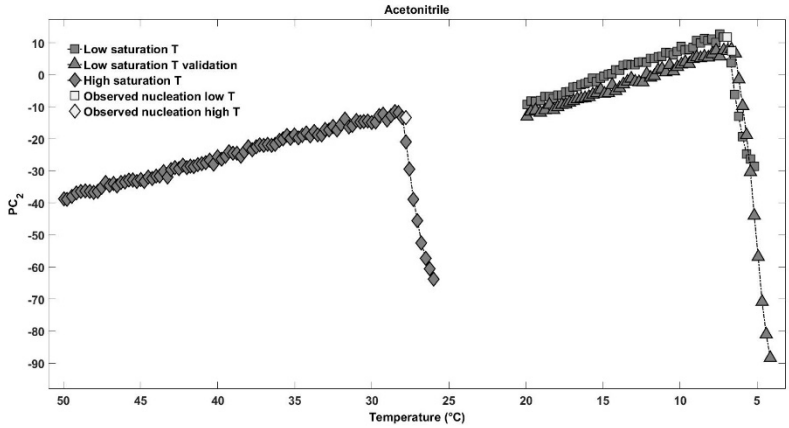

Figure 25 – Second score of INA in acetonitrile compared with visually observed nucleation temperature: training set (10°C) and validation set (10 and 35°C)

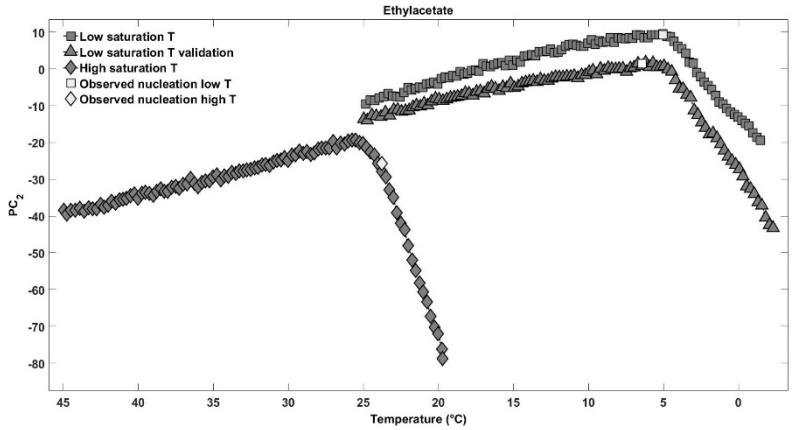

Figure 26 – Second score of INA in ethyl acetate compared with visually observed nucleation temperature: training set (10°C) and validation set (10 and 35°C).

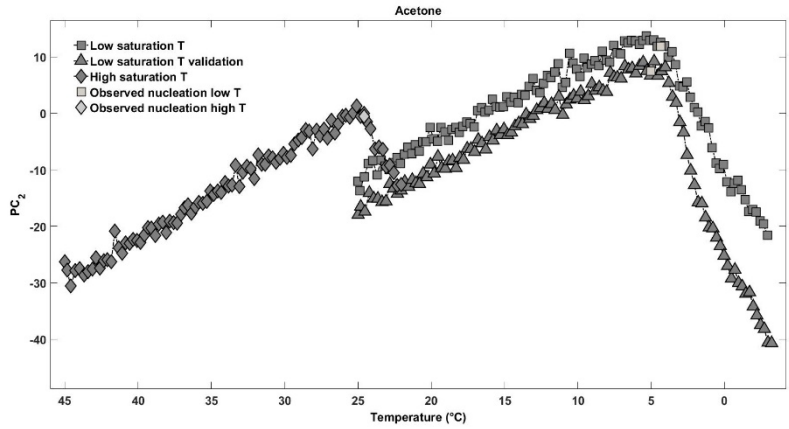

Figure 27 – Second score of INA in acetone compared with visually observed nucleation temperature: training set (10°C) and validation set (10 and 35°C)

5.3.2 On-line detection of nucleation

In this section, algorithms for the analysis of transient data were developed and tested. The intensity of the infrared spectra collected during crystallization are influenced by the change of INA concentration but also by the temperature since it affects the vibration intensity and frequency of molecular bonds (as investigated by Simone et *al.*, 2014, Cozzolino et *al.*, 2007). The in-line implementation of these algorithms aims to detect nucleation as spectra are collected. Particularly, the MWPCA was implemented and the results were compared with those obtained through static PCA.

5.3.2.1 Static PCA for on-line monitoring

The static PCA was firstly implemented to show its ability to detect faults. For its implementation, spectra were baseline corrected with a linear interpolation function, algorithm developed by Hrovat (2009) and pre-processed through mean-centering and standardization.
The static PCA was built considering a training set of spectra collected at T = [24.9 ÷ 13.61] °C that is higher than the saturation temperature. While, the prediction set consisted of the observations collected at temperature from 13.6 °C to -9.78 °C. The first component explained 75.75% of the variance, the second 9 % and the third 3.2% while each of

the subsequent components added only about 1%. Therefore, three principal components collected most of the variance present in the evolving system. T^2 and Q statistics were estimated for both training and prediction set (depicted as open squares and grey circles, respectively) and represented in Figure 28. As it was expected, the values of the statistics for the training set fell inside the normal operating region, only the Q statistic pertaining two observations was slightly higher than the limit. However, their value was quite close to the limit one and they can be classified as false positive. Incidentally, one should note that the false positive ratio in the training set was acceptable (3% of the total observations) and it was comparable to the significance level chosen for the limit value of the statistics (α=5%). Concerning the prediction set, as reported in Pöllänen et *al.* (2006), if more than three observations exceed the limit, the system can be considered to approach the nucleation. However, it can be observed that the values of the two statistical indexes are clearly non-stationary (as also reported in Ku et *al.*, 1995) and they increase with temperature. This feature was due to the fact that a static PCA was used to describe temperature-dependent data. In fact, the control charts show that the data collected for temperature below 7 °C result classified as out-of-control although nucleation has not occurred yet. Therefore, considering the visually observed nucleation point (T_{NUCL}= -7.38 °C), the Q statistic correctly classified 30.85 % of the prediction set and the false positive rate is 77.4 %. On the base of the T^2 chart, 37.2 % of spectra were correctly classified and the false positive rate is 70.23 %. Moreover, the alarm triggers at temperature T= 6.62 °C for T^2 and T=7.36 °C according to Q chart, that is about 14 °C or, alternatively 30 min, earlier than the onset of the crystallization. This result was due to the fact that the drift of the spectral variables during the crystallization caused the change of the mean and covariance. As a consequence, the false positive rate was quite high and the control charts could be misleading. Although the change of slope of the statistical indexes values could give a definitive indication of the beginning of the nucleation, static PCA led to questionable outcomes, since the drift was intrinsic and was not to be considered as a fault in this case study. Thus, a more reliable procedure would be preferable during monitoring of crystallization to detect the nucleation and avoid both false positive and negative.

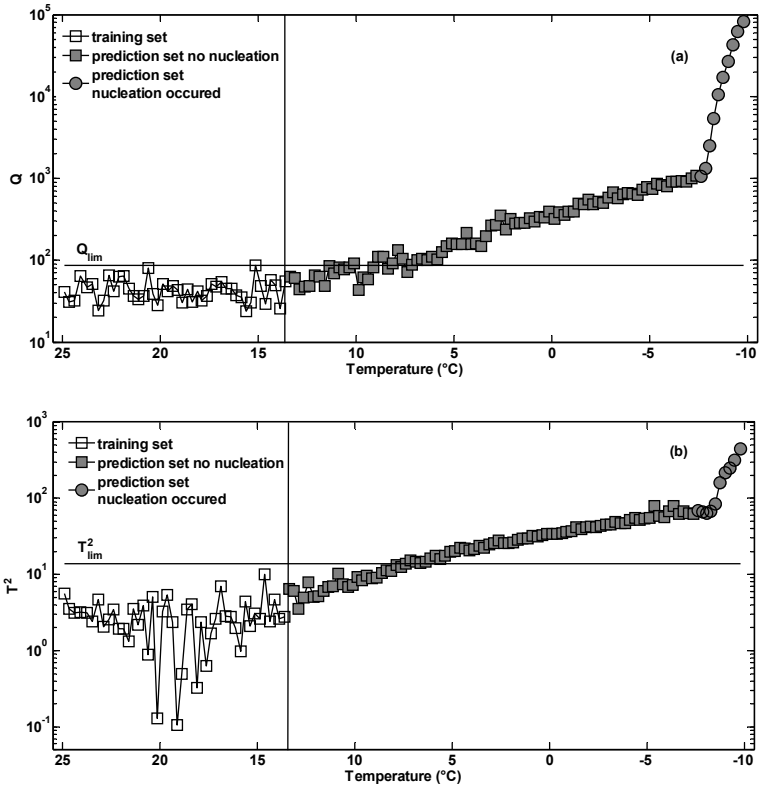

Figure 28 - Statistical monitoring using static PCA applied to Infrared data of INA in methanol. (a) Q statistic; (b) T^2 statistic. The values estimated for the prediction set show a dependence on the temperature, the false positive rate is 77.4 % and 70.23 % for Q and T^2, respectively.

Moreover, another important aspect to take into account when a dynamic process is modeled through static PCA, is the selection of the proper training set. Indeed, the effectiveness of the statistical control depends on the size of the window, that is on the number of the observations included in the training set. In order to test the influence of the window size on the statistical control, static PCA was implemented considering a training set whose size was increased from the beginning of the cooling T=24.9 °C to a value near the saturation temperature. Three latent variables were chosen that explained an average cumulative variance of 85.4 %. Concerning the prediction set, it consisted of the observations collected at temperature T= T_{train} to the end of the experiment. Figure 29a shows that, when the size of the training set was increased from 20 to 40 observations

(that means from the temperature range T= 24.9-19.86 °C to T= 24.9 to 13.61 °C), the threshold temperature denoting the approach nucleation decreased from about 17 to 7 °C. This means that, if the training set size is too low and the conditions (temperature) at which the observations are collected are quite far from the actual nucleation temperature, the alarm could trigger too early. In fact, comparing this results with the visually observed nucleation point (T_{NUCL}= -7.38 °C), the alarm would start from 24 °C to 14 °C before the nucleation occurrence. Figure 29b depicts the false positive rate detected from the T^2 and Q control charts depending on the size of the training set. One should notice that the ratio is quite high and does not decrease significantly with the window size (from 92% to 77%).

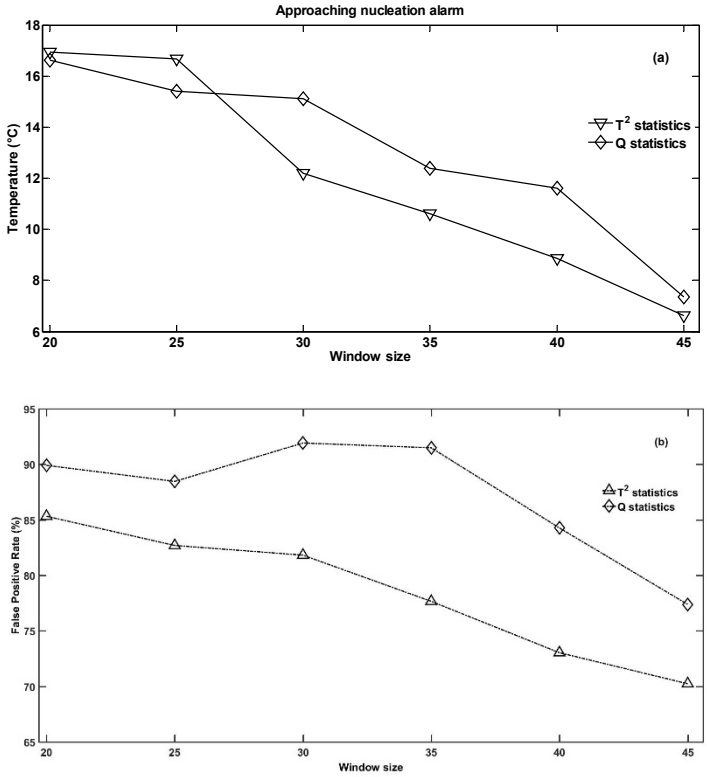

Figure 29 – Results after implementation of static PCA increasing the training set from a size of 20 (T= 24.9 to 19.86°C) to 45 (T= 24.9 to 13.61 °C). (a) The temperature at which the alarm should trigger to indicate the approaching nucleation was detected from the T^2 and Q control charts: when the size increased, the alarm temperature decreased. (b) the false positive rate calculated for T^2 and Q, decreased as the size increased.

5.3.2.2 Moving Window PCA

In this case study the external variable is the temperature. Before the implementation of the procedures illustrated in section 5.2.2, spectra were baseline corrected. Spectra pre-processing through mean-centering and standardization, was performed each time the window was moved. In principle, the system should fall into the out-of-control state at the onset of crystallization. To choose the window size L, the following parameter were taken into account: the minimum number of points collected in the experimental window was L_{min}= 20, therefore the first window included the 20 observations collected between T= 24.91 and 19.86 °C; whereas the maximum size L_{max}=45, the window included 45 observation collected from T= 24.91 up to T= 13.61 °C, value near the saturation temperature, since for temperature lower than the saturation one, the nucleation may occur. The PCA model was developed with 3 principal components that explained between 82 and 88.9% of cumulative variance (depending on the window size). While, for the validation set, three subsequent observations (V=l+1, l+2, l+3) were considered. The SSPE calculated through Equation (26) is shown in Figure 30: as it can be seen, the minimum size of the moving window was estimated as equal to 43 observations, corresponding to ΔT=10.8 °C.

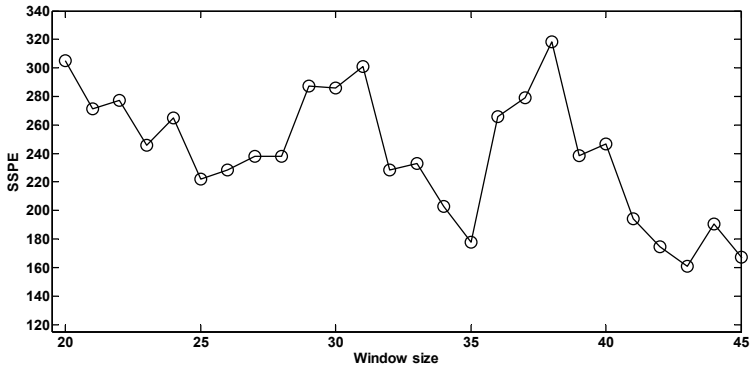

Figure 30 – SSPE calculated varying the window size. The minimum value was observed for L=43.

MWPCA was implemented according to the procedure shown in Figure 20: the PCA model was built selecting three latent variables explaining an average cumulative variance of 86.66% and the results of the on-line monitoring procedure achieved for the most representative windows were

depicted in Figure 31 to Figure 33. T^2 and Q control charts are represented in Figures (a) and (b), while Figure (c) shows the contribution plot where the contribution of each wavenumber to the statistics was reported, allowing the detection of the spectral region significantly changing due to nucleation.

Figure 31 depicts the results obtained from MWPCA built considering a training set where observations collected at T = 24.91÷13.61 °C were included. In more detail, Figure 31a and Figure 31b, show the values of T^2 and Q, they are indicated by open squares for the training set and grey circles for the prediction set, while the control limits were evaluated as in Equations (13) and (15) at 95% confidence level. Figure 31c shows the contribution plot, reporting the standardized residuals with respect to the corresponding wavenumbers as they were computed according to Equation (18). For sake of completeness, the main INA peaks are represented as open triangles. The threshold value for the contribution $\mathbf{c}^Q$, reported with the dotted line, was computed with a significance value α = 5% and modified according to Bonferroni adjustment. As it can be observed, (Figure 31a), the three subsequent observations, reported as grey circles, could be classified as in-control, and the contribution of each spectral variable did not exceed the control limit.

Concerning the results achieved for the window between T=2.86 and -7.63 °C (Figure 32), the prediction set exceeded the limit value of Q statistic (Figure 32b) and the onset of the crystallization could be detected at T= -7.86 °C, value quite near (-0.48 °C) to the visually observed one (T= -7.38 °C). Indeed, it could be inferred from the contribution plot that peaks related to INA located at 1689 and 1451 cm^{-1} were increasing and exceeded the threshold value. On the other hand, the T^2 values could be classified as in-control, since the deviation from the model was limited and T^2 was not probably able to detect it yet.

Since three observations were greater than the Q threshold, the window was not moved anymore and new observations were added to the prediction set. Figure 33a clearly shows that the values of Q statistic for the prediction set were greater than the threshold and were considerably deviating. Moreover, it could be inferred from the contribution plot how the contributions corresponding to the INA peaks dramatically increased, while the solvent ones did not change. On the other hand, according to the T^2 chart, nucleation should be detected at T= -8.75 °C, this delay could be

due to the fact that in general the T^2 is less sensitive to deviations than Q (Qin, 2003).

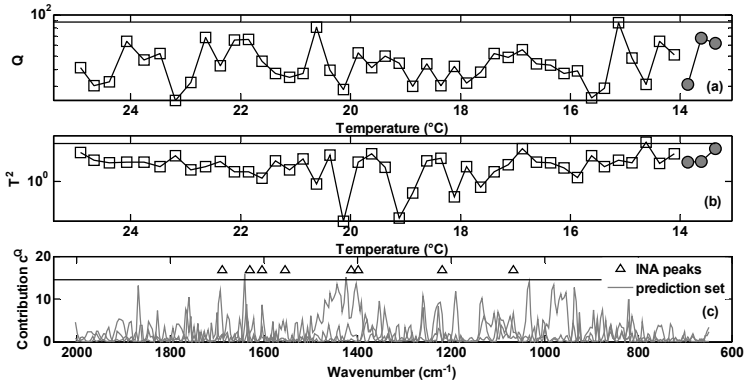

Figure 31 – Statistical monitoring of cooling crystallization of INA in methanol through Moving Window PCA built considering a training set (open squares) of observations collected at T=24.91÷13.61 °C. T^2, Q and $\mathbf{c}^Q$ control charts are reported, dash-dotted lines represent the control limits. The prediction set (grey circles) was classified as in-control, then no nucleation had occurred yet.

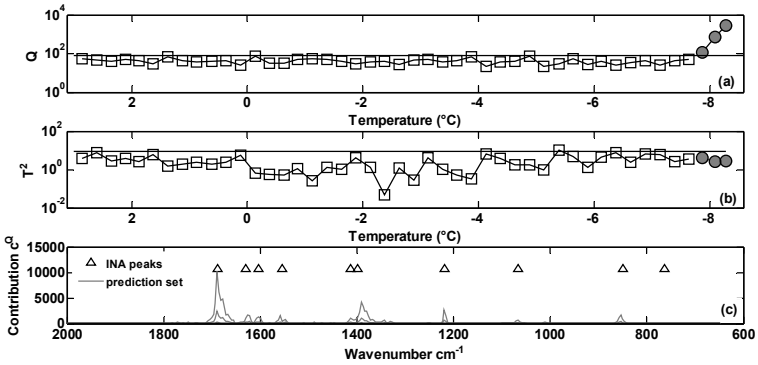

Figure 32 – Statistical monitoring of cooling crystallization of INA in methanol through Moving Window PCA built considering a training set (open squares) of observations collected at T=2.86÷ -7.63 °C. T^2, Q and $\mathbf{c}^Q$ control charts are reported, dash-dotted lines represent the control limits. The prediction set (grey circles) was classified as out-of-control according to Q control chart, then the onset of the crystallization was detected at T= -7.86 °C.

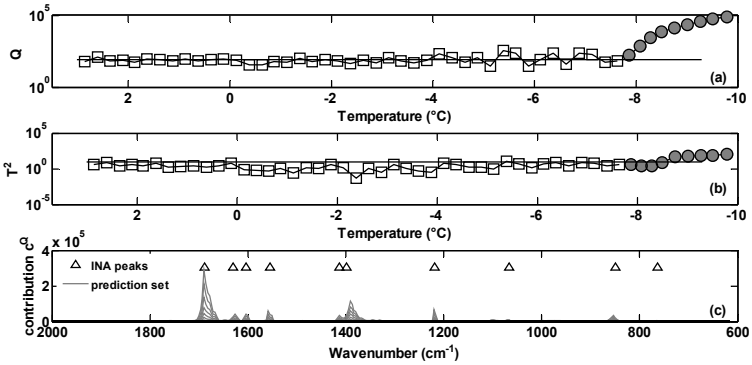

Figure 33 - Statistical monitoring of cooling crystallization of INA in methanol through Moving Window PCA built considering a training set (open squares) of observations collected at T= 2.86 ÷ -7.63 °C. T^2, Q and $\mathbf{c}^Q$ control charts are reported, dash-dotted lines represent the control limits. It can be noted that the future observations were strongly deviating from the NOC and the crystallization was continuing.

A magnification of the contribution plot estimated (in Figure 33c) for the observations collected after the nucleation is reported in Figure 34. At the beginning the extent of the contribution was limited, while it increased as crystallization progressed from about 10 (before the nucleation, see Figure 33a) to 10^4 (after the nucleation). During the implementation of the MWPCA only six false positives could be globally detected, while false negatives are zero for Q statistic, therefore the global false positive rate is 5.8%. Regarding T^2, seven false positives were observed and the global false positive rate was 6.7 %, value higher than Q statistic. It should be noted that also four false negatives were present, since nucleation was detected later than Q statistic, leading global false negative rate to 30.7 %. This high value reflected the ineffective monitoring performances of T^2, at least for the case at hand. Indeed, as explained in section 2.5.1, spectral variables should significantly deviate from the origin of the new subspace so that is detected using T^2 (Qin, 2003). Therefore, Q statistic demonstrated more able to discriminate out-of-control observations and detect the nucleation earlier than the T^2.

The MWPCA reveals as a suitable method to monitor transient processes like crystallization, since it demonstrates to detect nucleation more accurately than static PCA, indeed, the false positive rate has noticeably decreased from 77.4% (static PCA) to 5.8% (MWPCA), improving then the reliability of the control system.

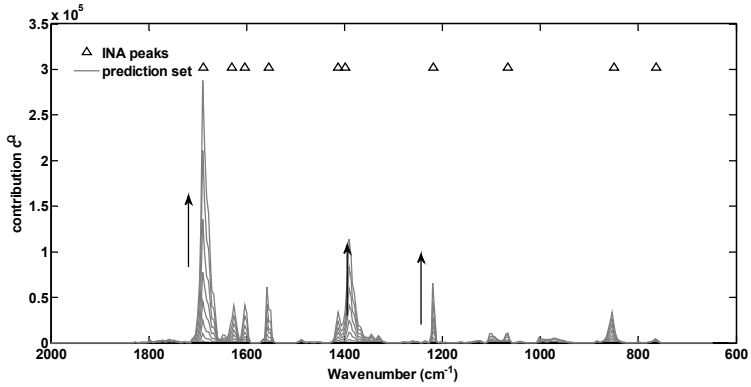

Figure 34 – Contribution plot of the observation detected as out-of-control. Contribution related to INA peaks were increasing, while solvent ones did not vary. Since the nucleation occurred the INA peaks were supposed to change.

Final remarks

Summarizing, the implementation of static PCA on IR data demonstrated that the nucleation temperature could be easily detected even when API concentration was different from the data used to create the model. However, the on-line detection of the nucleation required a methodology able to model dynamic data such Moving Window PCA proposed in this thesis. As a result, it demonstrated capable of correctly identifying the out-of-control status, otherwise not achievable through static PCA. Indeed, T^2 and Q control charts based on MWPCA detected the onset of crystallization (T= -7.86 °C). The estimated value was quite close to the visually observed one (T= -7.38 °C). In addition, INA peaks were identified as the mostly contributing variables to the out-of-control status through contribution plots. Finally, it is worth noting that it improved the control system reliability since the false positive rate had noticeably decreased from 77 % (static PCA) to 5.8% (MWPCA). However, in order to further improve the accuracy of the detection, it would be interesting to investigate other approaches such as the combination of Recursive PCA and MWPCA (Jeng, 2010) and compare the results.

Chapter 6

Dissolution of dish paste

In this thesis, the dissolution of a dish paste was investigated. It is in gel-like phase and consists of two surfactant pastes where water and surfactant A and B are present, respectively. Further details on the nature of the chemicals cannot be reported since they are covered by Confidence Disclosure Agreement with the P&G. During production of liquid dish detergent, a dish paste is dissolved in solvent, usually water. The diffusion of the surfactants paste into the solvent, may lead to the formation of mesophases at the interface that can influence how the system evolves during dissolution (Gradzielski, 2003). For this reason, the dissolution of the dish paste represents a key factor to interpret paste dissolution experiments, to properly design dissolution process and improve liquid dish detergent production, but also during its use at home. Surfactants dissolution is an interesting topic that is not often investigated, maybe due its complexity and the scarcity of systems where the equilibrium phase behavior can be easily explained (Warren & Buchanan, 2001). In general, surfactants dissolution is examined through optical microscopy (Chen at *al.*, 2001), where the radius of a surfactant drop is tracked during dissolution, but no chemical information about the species and the morphological changes can be inferred. To this aim, hyperspectral techniques have been recently revealed very useful, since they allow to infer chemically and spatially resolved information.

The employment of Hyperspectral Imaging techniques for quality monitoring has significantly grown over the last decades. It begun in the early 70's with mainly applications to remote sensing (van der Meer et *al.*, 2012; Lee et *al.*, 2011), then it started to extend to different research areas such as food sciences (Cheng et *al.*, 2016), pharmaceutical research (Terra et *al.*, 2014; Alexandrino et *al.*, 2015), but also in cultural heritage (Sciutto et *al.*, 2012), in wood quality control (Burud et *al.*, 2014) and polymer research (Mukherjee et *al.*, 2015) for detection of defect in the surface. Hyperspectral Imaging comes from the combination of image analysis and bulk spectroscopy (e.g. UV, Infrared, Raman spectroscopy). The advantage of this recent technique with the respect to image analysis and bulk spectroscopy is that information about the distribution of the components can be gathered.

Traditional image analysis is based on the fact that the optical properties of a product (food, surface, drugs) provide spatially resolved information about the quality of the product such as defects, structural changes and

texture features (Prats-Montalbán et *al.*, 2011). The acquired images consist of a single measurement (in gray images) or three color-related values (in RGB images) for each pixel. Time Lapse Photography represents a particular kind of image analysis, where the images are collected over the time to detect and monitor changes of optical properties in the sample occurring. In this case, the information is not only spatially but also time resolved. It is employed e. g., in remote sensing (Nagai et *al.*, 2016), in *in vitro* fertilization to monitor the embryo growth (Mandawala et *al.*, 2016). Nevertheless, traditional image analysis does not provide any chemical information about the composition of the samples.

On the other hand, bulk spectroscopy (Raman, UV, IR, etc.) is able to give average chemical information about the sample being analyzed and its composition. It has become popular in industry for quality monitoring since it is a non-invasive and non-destructive technique. In this framework, an evolution of the bulk spectroscopy is represented by *in situ* spectroscopy where measurements are collected over time and helps in process understanding and reaction monitoring (e.g., Leineweber & Mittemeijer, 2012; Wolf et *al.*, 1999). Despite the great potentialities of bulk spectroscopy, its use is limited only to systems where there is no spatial distribution of the components, since no spatially resolved chemical information can be gathered.

Therefore, the hyperspectral imaging technique represents a valuable tool that allows to overcome the issues related to the image analysis and bulk spectroscopy. Briefly, the hyperspectral imaging technique consists of acquiring images of the sample (in the most general case also over the time), a spectrum is collected for each point of the 1D, 2D or 3D mesh of the image. An illustrative example of the measurements performed by this technique is shown in Figure 35, where phenomena occurring in a sample are investigated. At the beginning of the experiment the sample appears homogeneous. Indeed, the spectra (Raman, UV, Infrared) collected along the space are very similar. After a certain time, formation of a new component due to reaction, degradation or separation of components can be detected in terms of appearance of another peak arising in the spectrum collected in the upper part of the sample.

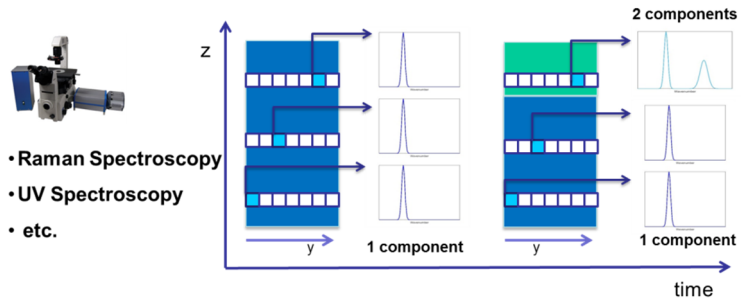

Figure 35 – Principles of Hyperspectral Imaging Technique.

Among the different hyperspectral analytical technique, Confocal Raman microscopy represents a powerful tool (Everall, 2009), that allows to gather spectral information at specific position in the sample being analyzed. Moreover, Raman spectroscopy are weakly influenced by the presence of water in the sample at least when compared to other analytical techniques, such as, e.g., infrared spectroscopy.
As mentioned previously, the aim of the case study is to investigate the dissolution of a dish paste. To the authors' knowledge, no works present in literature study dissolution of surfactants paste at microscopic level, but only by means of optical microscope. Therefore, Confocal Raman microscopy was explored to assess whether it can be a useful tool to study the dissolution of paste A and B.

6.1 Experimental

6.1.1 Materials

The dish paste consisted of paste A and B which contain water and surfactant A and B, respectively. The ratio of the activity of surfactant B and A in the dish paste was 4.4.

6.1.2 Experimental set-up

The experiment was performed at the Procter & Gamble R&D Brussels Innovation Center using a confocal Raman microscope. It was an XPloRA (Horiba) equipped with multiple objectives allowing spatial resolution down to 1 micron (3-4 micron in Z direction). Multiple laser wavelengths

are available: 532 nm, 638 nm and 785 nm. The spectra cover the range from 2561 to 287.4 cm^{-1} with an average wavenumber resolution equal to 5.6 cm^{-1}. The spectrometer is also equipped with a glass capillary device provided by the Imperial London College (Figure 36), whose dimensions are 1 mm width and 2 cm length and it is provided with two open edges, where the height of the indent is 0.11 mm. The dish paste was introduced inside the capillary from the inlet open edge, then the capillary was located inside a water bath to allow the water to flow through the capillary and wet the dish paste during the experiment. Then, Raman spectra were collected from the open edge (indicated as z_0) to 5 mm with a spatial resolution of 0.5 mm, from the beginning of the experiment until 85 min every 5 min. Moreover, an additional measurement was carried out at 135 min.

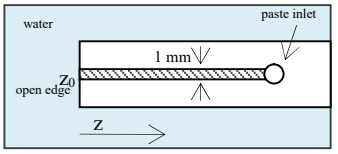

Figure 36 – Capillary device where the dish paste was introduced and it was used in the Raman confocal microscope to carry out the dissolution experiment.

6.1.3 Raman spectra

Raman spectra of the surfactants pastes without the capillary and pure water in the capillary were collected before starting the experiment and they are represented in Figure 37a. As it can be observed, the spectra of the paste A and B are quite similar and many peaks are located at the same position. However, some differences can be recognized, in detail, peaks located at 3027, 952.5 and 761.9 cm^{-1} pertain to the only spectrum of paste A, whereas surfactant B has only one distinguishing peak at 825.9 cm^{-1}. Spectrum of the capillary is also reported, it can be noted that it shows peaks at 1089 cm^{-1} and a smaller contribution at 546.9 cm^{-1}, moreover a drift of the baseline (due to the fluorescence effect) is observed at higher wavenumbers. Spectrum of pure water without capillary is characterized by peaks at 448. 3 and 794 cm^{-1}. Although the scattering coming from the water is not directly comparable with the pastes scattering, it can be observed that the contribution of the water in the pastes A, paste B and water with the capillary spectra seems to be

negligible. In Figure 37b, the Raman spectrum of the dish paste is reported. Due to the glass, the signal is amplified, the baseline shows an offset and a drift at higher wavenumbers. In addition, peaks of surfactant A and B partially overlap in the spectral range between 700 and 900 cm^{-1} (see magnification of Figure 37b).

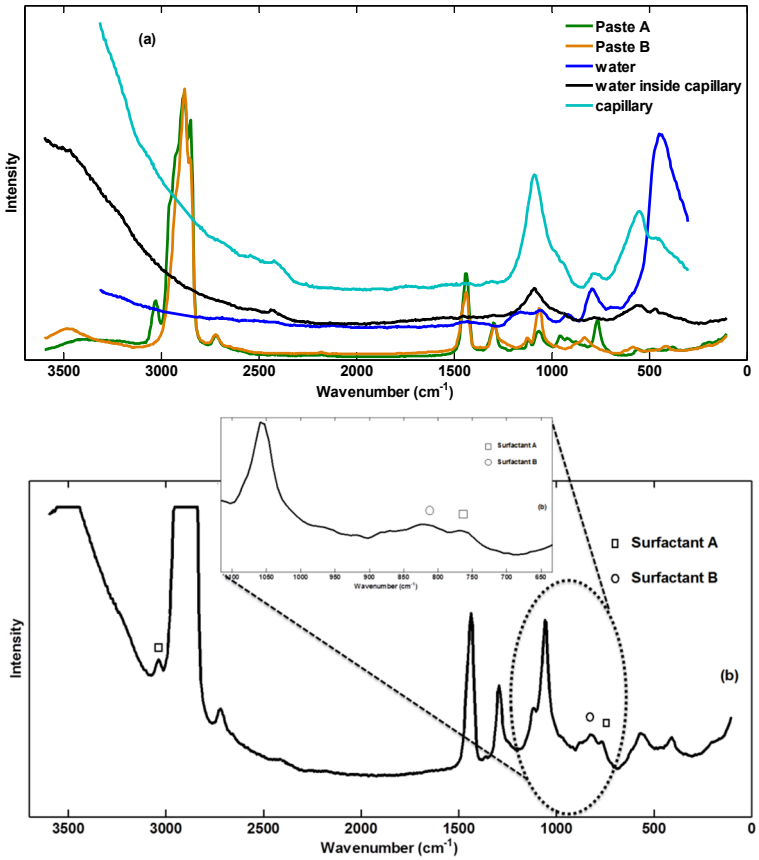

Figure 37- (a) Raman spectrum of paste A, paste B, water without capillary, water in capillary and capillary. (b) Spectrum of the mixture of paste A and B in the capillary.

6.2 Methods

A hyperspectral image consists of spectra collected for each point of the 1D, 2D or 3D mesh of the image of the sample. As a result, when spectra are collected also along one spatial variable z and over the time, hyperspectral data can be arranged in a 3-way matrix $\underline{\mathbf{X}}_{(I\times J\times K)}$ and can be

represented as a hypercube, where each row represents the spectrum collected along at *J* spectral variables and along *I* spatial variable and over *K* sampling times (blue cube in Figure 38). Therefore, chemometric tools represent a valuable tool to extract information, monitor the quality, study the compounds distribution in the sample. Principal Component Analysis is one of the most common method in hyperspectral data analysis, it is used to detect and monitor defect in samples and for Statistical Process Control (SPC) (Prats-Montalbán et *al.*, 2011) in Multivariate Image Analysis (MIA). It reduces the dimension of the data sets and decomposes the spectral data calculating few latent variables that are able to describe main changes and features, where each latent variable can be associated to a pseudo-spectrum. However, they do not always have a clear physical meaning since they will result as a linear combination of the actual spectra. To overcome this issue, Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is preferable (de Juan et *al.*, 2014), since it can estimate the distribution maps and pure spectra related to the image constituents of a sample, that are more easily interpretable.

Since multivariate techniques perform the decomposition on a two-way matrix, hyperspectral data need to be unfolded beforehand. There are different approaches used for the unfolding, the most commons ones (Camacho et *al.*, 2008) are the row wise where the hypercube is rearrange in the matrix $\mathbf{X}_{(I\times JK)}$ as depicted in Figure 38a, whereas the other approach arranges the matrix $\underline{\mathbf{X}}_{(I\times J\times K)}$ into the matrix $\mathbf{X}_{(IK\times J)}$ as represented in Figure 38b. Since, Multivariate Curve Resolution was employed in this thesis, the column-wise unfolding procedure was carried out (de Juan et *al.*, 2009). The procedure used for the analysis of hyperspectral data is summarized in Figure 39.

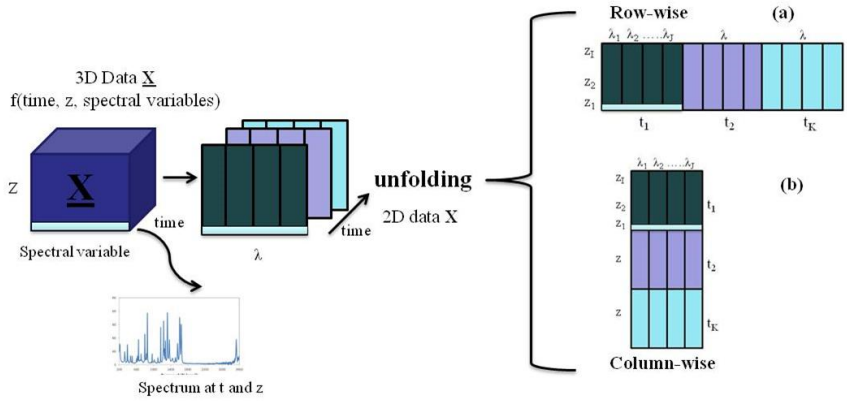

Figure 38 – Unfolding procedures of the 3-way matrix X

The decomposition performed by MCR reported in Equation (43), aims to infer the concentration profiles and spectra of the species.

$$\underset{(IK\times J)}{\mathbf{X}} = \underset{(IK\times A)}{\mathbf{T_A}} \cdot \underset{(A\times J)}{\mathbf{P_A^T}} + \underset{(IK\times J)}{\mathbf{E}} \quad (43)$$

In hyperspectral imaging, constraints usually used for concentration profiles, like unimodality, closure or hard-modeling results not always suitable due to the discontinuity introduced with the unfolding procedure (de Juan et *al.*, 2014). Therefore, only non-negativity is applied to concentration profiles of the different components. Another constraint to be taken into account is that estimated spectra should be non-negative. Concerning the initial guesses required to implement the algorithm, a preliminary estimation of the concentration along the space and/over the time is not feasible through methods like Evolving Factor Analysis, since unfolding procedure introduces discontinuities in the unfolded data. Therefore, it would be preferable to use methods that aim to estimate the 'purest' spectra from the hyperspectral data matrix, such as SIMPLISMA (Winding et *al.*, 1997). On the other hand, if the experimental spectra of pure components are available, they can be considered as initial guesses as well. Once the decomposition is performed, the concentration matrix $\mathbf{T}_{A(IK\times A)}$ needs to be reshaped in the hypercube $\underline{\mathbf{T}}_{A\ (I\times A\times K)}$, similarly to the procedure followed for the unfolding of the experimental hypercube. Spectra and concentration images can be then represented and analyzed.

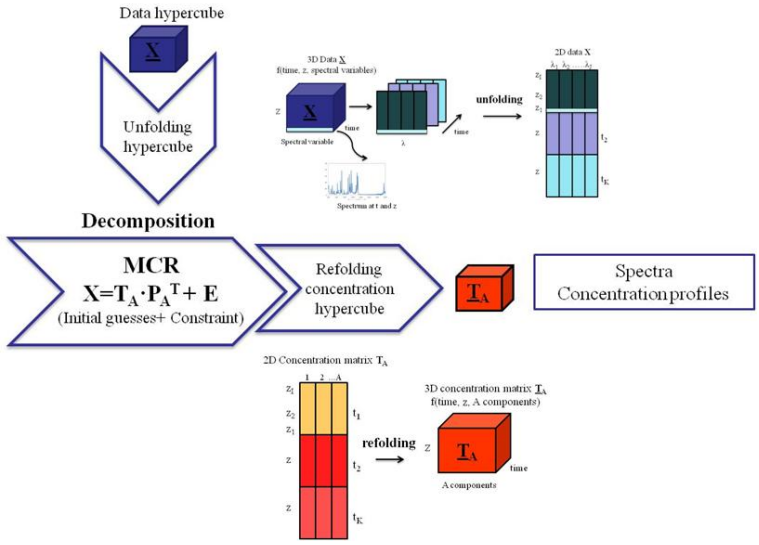

Figure 39 – Flowchart of the procedure used for hyperspectral imaging analysis.

6.3 Results

Since the characteristic peaks of the two pastes are located between 902.1 and 691 cm^{-1}, only this spectral range was considered for the multivariate analysis. Then, Raman hyperspectra were arranged in the 3-way matrix $\underline{\mathbf{X}}_{(11\times20\times34)}$, that was unfolded into the matrix $\mathbf{X}_{(220\times34)}$. Four different components were considered as initial guesses for the implementation of MCR-ALS algorithm: the experimental Raman spectra of paste A, paste B, empty capillary and water (Figure 37a). The spectrum of the capillary is taken into account as well, since it has been noted that the contribution of the glass to the spectra may change and influence the intensity of the collected spectra during the experiment. Non-negativity constraint was applied to both concentration profiles and spectra. Spectra estimated through MCR-ALS are represented in Figure 40, where they are compared with the experimental ones.

As it can be seen, spectra of component 1 to 3 correspond to experimental spectra of paste A (Figure 40(i)), paste B (Figure 40(ii)) and capillary (Figure 40(iii)), respectively. Estimated water spectrum results to be zero, thus MCR-ALS did not appear to be able to distinguish water contribution from paste A and B, probably because negligible. The evaluation of the

spectra by means of MCR was quite accurate, indeed the Pearson correlation coefficient between the estimated and the experimental spectra (see Equation (35)), was equal to 0.8511, 0.9474, 0.9555 for spectra of paste A, paste B and glass, respectively.

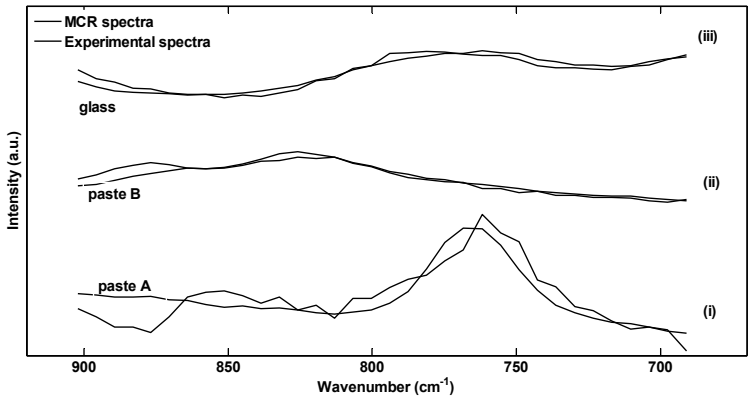

Figure 40 - Spectra estimated through MCR-ALS compared with the experimental spectra of paste A (i), paste B (ii) and capillary (iii). Estimated water spectrum was equal to zero, since probably its contribution was negligible and not distinguishable from the other components spectra.

The concentration profiles were normalized with respect to maximum value assumed along the space and over the time for each concentration profile. In Figure 41a the normalized concentration of paste A is represented in the *z-t* plane as a surface, two different 2D plot could be extracted from this 3D surface in order to focus more on the variation of the concentration over the time or along the space. the concentration over the time at different position in the capillary is reported in Figure 41b: as it was expected, the concentration estimated at each position (therefore the dissolution) decreased over the time, but it slowed down approaching the paste inlet (z = 5 mm). It could be also appreciated that the concentration profile at z > 1.5 mm showed an initial lag period, then an exponential decrease. On the other hand, Figure 41c highlights the concentration profiles along the capillary estimated at different times: the concentration increased along the capillary at each time. It is worthwhile noting that concentration reached a constant value at z = 5 mm, implying that dissolution did not occur at z = 5 mm. Similar consideration could be carried out for the concentration profiles of paste B depicted in Figure 42.

These concentration profiles represented the starting point to investigate and model the dissolution of paste A and B. It is interesting to note that the glass contribution to the spectra (depicted in Figure 43) seemed to increase where the presence of water was higher and the dish paste was dissolving. Therefore, the presence of interfering components that could influence Raman scattering over the time and along the space could be to take into account through MCR-ALS.

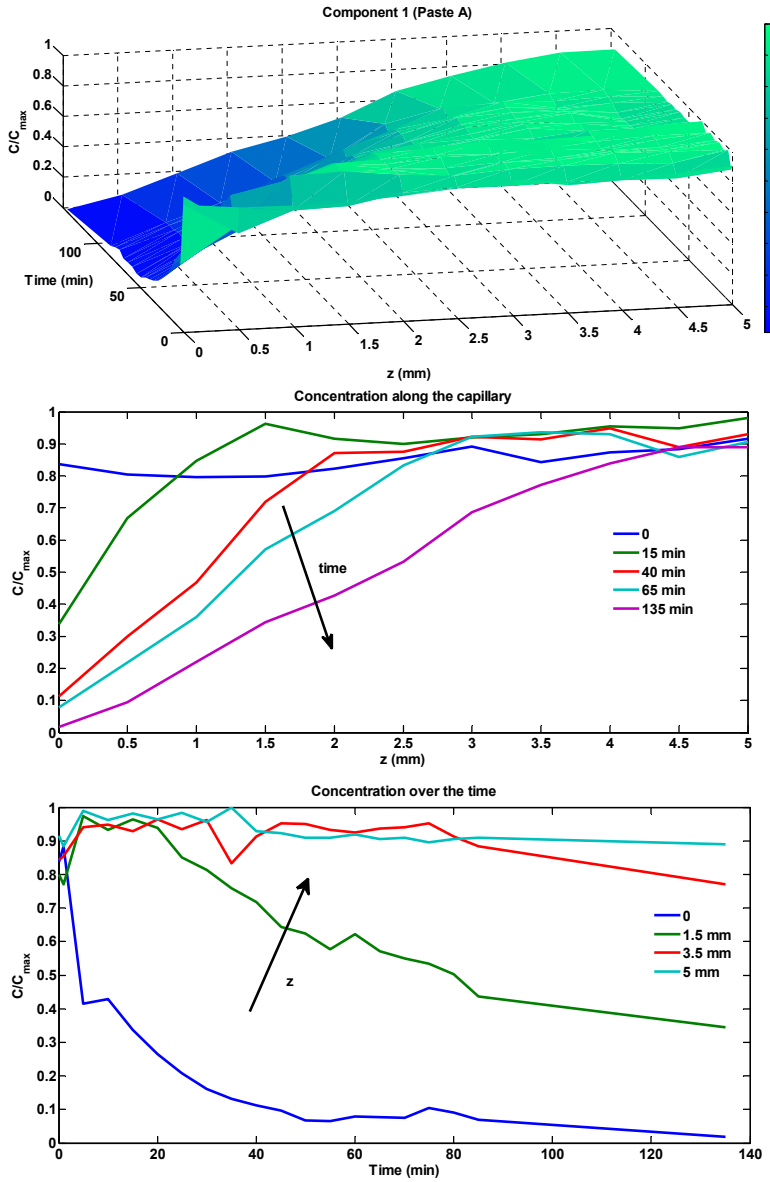

Figure 41 - (a) Concentration profiles of paste A. 2D graphs report the concentration profiles (b) along the capillary and (c) over the time.

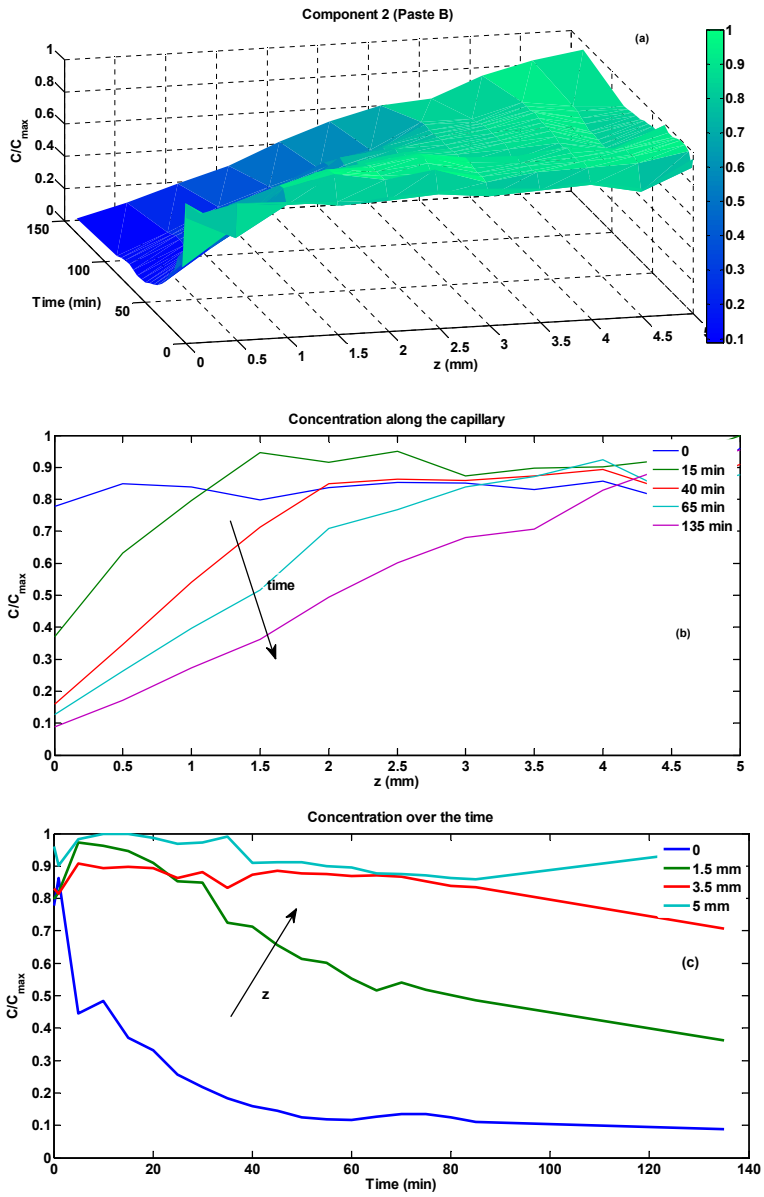

Figure 42 - (a) Concentration profiles of paste B. 2D graphs report the concentration profiles (b) along the space variable and (c) over the time.

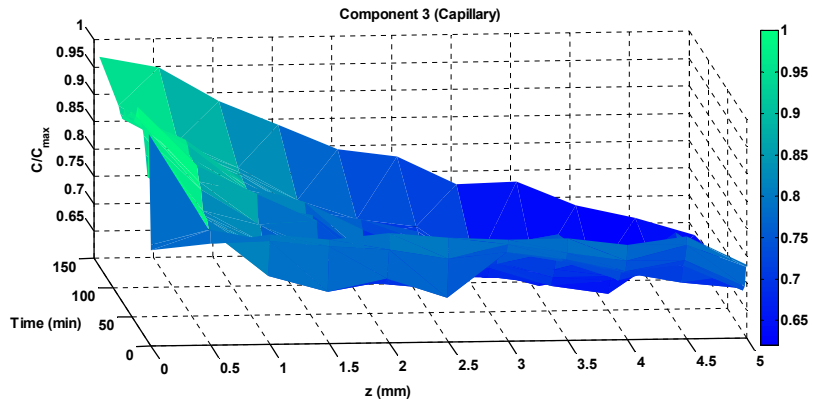

Figure 43 - Relative contribution of glass to the spectra over the time and along the capillary.

In order to assess whether the paste A and B dissolved similarly, as an illustrative example, the normalized concentration profiles of paste A and B estimated at z = 1.5 mm was considered. Moreover, the normalized height of peaks at 761.9 cm^{-1} (belonging to paste A) and 825.9 cm^{-1} (belonging to paste B) was reported in Figure 44b. As it can be inferred from Figure 44a, the profiles of paste A and B overlap fairly well, implying that two surfactants behaved in a similar fashion. This was also confirmed by the normalized peaks height curves shown in Figure 44b that decreased together as well.

Different empirical models are available in literature to describe dissolution over the time (Costa & Lobo, 2001). Here, the dissolution rate d of paste A and B was estimated using a simple exponential decay model as expressed in Equation (44).

$$\hat{t}_z = b \cdot exp(-d \cdot t) + c \qquad (44)$$

The parameters of the exponential model were estimated for the normalized curves obtained by means of MCR-ALS (Figure 44a) and for the normalized peak height tracked over the time (Figure 44b). A non-linear regression was carried out through the Curve Fitting Toolbox of Matlab® which exploits the Levenberg-Marquardt algorithm to evaluate parameters. Results of the model calibration are summarized in Table 7 and the predicted curves are reported as dotted lines in Figure 44.

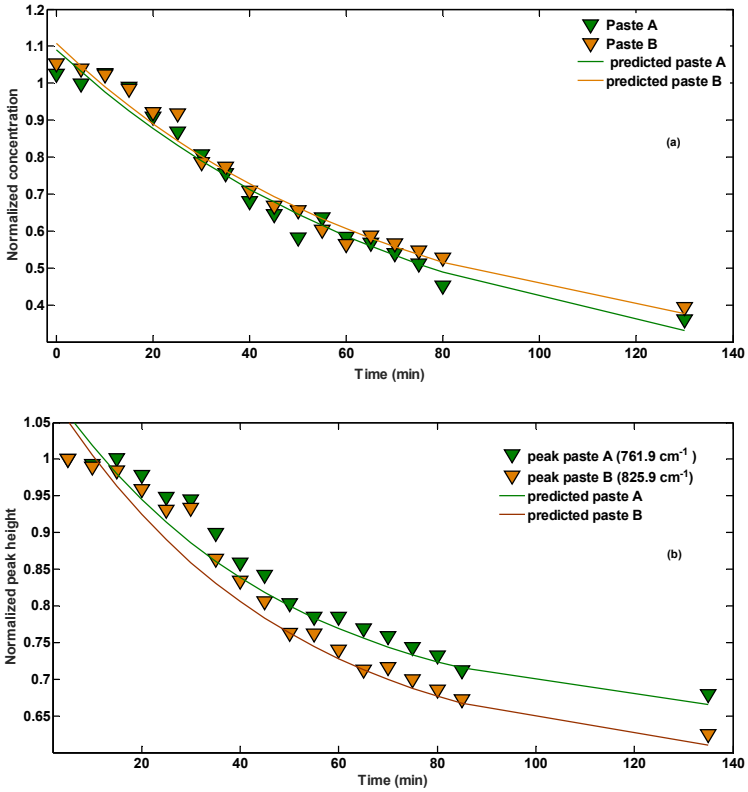

Figure 44 - (a) Comparison of normalized concentration profiles of paste A and paste B estimated through MCR-ALS at z=1.5 mm over the time. (c) Comparison of normalized peak height of paste A and paste B at z=1.5 mm over the time.

The model was able to describe both the MCR concentration profiles and the peak height dynamics. In the first case, the determination coefficient for normalized concentration profiles of paste A and B was equal to 0.969 and 0.976, respectively. As it regards the normalized peak height curves for paste A and B, a slightly higher determination coefficient was found, that was 0.986 and 0.9828, respectively. Note that the dissolution rate *d* between MCR concentration profiles of paste A and B was qualitatively similar, at least they have the same order of magnitude (0.013 and 0.01457 min^{-1}, respectively). The values of *d* obtained from the normalized peaks height between paste A and B were similar (0.02156 and 0.02115 min^{-1}, respectively). Although the dissolution rate estimated for the peak height was double the values of the MCR dissolution rate, the

results obtained could suggest that both paste A and B were dissolving at the same rate.

	Normalized concentration (MCR-ALS)		Normalized peak height	
	paste A	**paste B**	**paste A**	**paste B**
dissolution rate d (min^{-1})	0.013 (0.007209÷0.01885)	0.01457 (0.009273÷0.01986)	0.02156 (0.01631÷ 0.0268)	0.02115 (0.01542÷0.02688)
pre-exponential factor b	0.93 (0.7177÷ 1.143)	0.8652 (0.7136÷1.017)	0.47 (0.4349÷0.5051)	0.5263 (0.4824, 0.5702)
c	0.1608 (-0.0722÷0.3938)	0.247 (0.07947÷ 0.4144)	0.64 (0.5977÷0.6824)	0.58 (0.5265÷0.6336)
R^2	0.969	0.976	0.986	0.9828
RMSE	0.03974	0.03399	0.01284	0.01584

Table 7 – Results of non-linear regression of the exponential decay used to describe dissolution profiles. The confidence intervals are also reported.

In conclusion, hyperspectral confocal Raman microscopy in combination with Multivariate Curve Resolution demonstrated to be able to infer concentration of the pastes A and B over the time and along the capillary with quite good accuracy. It should be remarked that, unlike peak height tracking, the multivariate analysis carried out through MCR circumvented the numerical problems that may arise due to the presence of interfering contribution (the glass) in the spectra during the experiment and aided in hyperspectra interpretation.

Chapter 7
Setting reaction of cementing materials

The case study here investigated is the setting reaction of Magnesium potassium phosphate ceramics (MKPCs). They are chemically-bonded ceramics (Wagh, 2004) attractive for applications like waste encapsulation, bone repair, natural fibre composites. They form by fast reaction of MgO with potassium di-hydrogen phosphate (KDP) in solution. The detected crystalline product is K-struvite (MKP): $MgO + KH_2PO_4 + 5H_2O = MgKPO_4 \cdot 6H_2O$. Although several mechanisms for this reaction have been proposed (Soudée & Péra, 2000; Wagh & Jeong, 2003; Hall *et al.*, 1988), kinetic analysis was accomplished for the first time using *in situ* synchrotron powder diffraction (Viani *et al.,* 2016). The first step of the reaction was recognized to be the dissolution of MgO, and the best fit of the derived kinetic curves was obtained with a combination of first order model followed by a diffusion control one. Crystallization of MKP occurs later, with a first reaction step described by the Johnson-Mehl-Avrami–Erofe'ev-Kolmogorov equation (Brown et *al.*, 1980). These experimental evidences suggested that MgO dissolves quickly in the aqueous solution producing an intermediate amorphous phase. This thickening layer shifts the mechanism toward a diffusion control. The presence of an amorphous phase acting as the precursor of the crystalline MKP, was previously inferred observing that, after 30 min from the beginning of the reaction, and over long times, its amount decreases whereas that of crystalline MKP increases (Viani *et al.*, 2015, Viani & Gualtieri, 2014). However, amorphous development during the first minutes, and the direct relationship with the onset of MKP crystallization, have never been shown before. Therefore, since reaction mechanisms and physical phenomena are not completely understood, investigating the evolution of the system by means of complementary techniques has become essential for process understanding.

Time-resolved in-situ experiments are well suited for extracting information about reaction mechanism and activation energies (Leineweber & Mittemeijer, 2012). The advent of fast detectors and high intensity sources (i.e. synchrotron radiation and neutron sources), allowing for good time resolution, led to an explosive development of in-situ kinetic experiments employing powder diffraction. Fast collection rates, available at laboratory, as well as at neutron or synchrotron facilities, require the treatment of a huge amount of data. Commonly, in a time-resolved diffraction experiment of a transformation reaction, kinetic

parameters are obtained from the analysis of the contribution of the crystalline fractions of the reactants and/or products to the detected total scattered intensity, as a direct evidence of the respective fractions transformed. This can be done by considering the integrated area of selected peaks in the powder diffraction pattern (Gualtieri *et al.*, 2012; Solberg & Hansen, 2001; Cattaneo *et al.*, 2003; Kubo *et al.*, 2004), or alternatively, through the full profile Rietveld refinement of each data set (Allen & Evans, 2004; Allen *et al.*, 2003; Müller *et al.*, 2009). The procedures can be partly or completely automated, allowing for the analysis of hundreds of data sets in few minutes (Stinton & Evans, 2007). In practice, the conditions can be much less favourable, hindering the application of such fast procedures. Regardless of the approach adopted, one should bear in mind that the dynamic of the system and the experimental set-up may hide pitfalls that can lead to an erroneous interpretation of the process (Norby & Schwarz, 2008; Scarlett *et al.*, 2010). On the other hand, single peak integration is not free from potential sources of error. Since intensity information is extracted from a limited portion of the entire spectrum, the method is particularly sensitive to every effect that can selectively alter the diffracted intensities, these include preferred orientation and peak overlap with other phases in the system (that can also be intermediate phases). However, the integration of single diffraction peaks in a manual fashion is sometimes the only practicable choice, making data treatment a very time consuming task.

Diffraction techniques encounter limitations in the complete characterization and description of the reaction when amorphous fraction is present. Several reactions of materials of technological or scientific importance see the intervention of an amorphous or poorly crystalline component. A far from exhaustive list include amorphous calcium phosphates formed by living organisms and during synthesis of materials for biomedical applications (Dorozhkin, 2010), chemically-bonded ceramics for bioengineering and structural applications (Wagh, 2004), cements (Kurdowski, 2014), biogenic and synthetic amorphous calcium carbonate (Cartwright *et al.*, 2012), amorphous pharmaceutical solids (Yu, 2001). The amorphous phase can be a metastable precursor of the crystalline counterpart lasting few minutes or hours (Politi *et al.*, 2004; Bolze *et al.*, 2002) or a main, stable, product of the reaction (Steinke *et al.*, 1988; Richardson, 1999). Many efforts have been devoted to the

assessment of the amount of amorphous phases with X-ray powder diffraction (XRPD); of the methods developed for the quantification of powder mixtures (Bish & Howard, 1988, Bates *et al.*, 2006; Lutterotti *et al.*, 1992; O'Connor & Raven, 1988; Scarlett & Madsen, 2006), some can be applied in laboratory in-situ time-resolved experiments (Bergold *et al.*, 2013; Jansen *et al.*, 2011). However, the number of examples of the latter type is limited, and to extend the application of such methods to include a larger number of systems and experimental set ups, is not straightforward and sometimes not possible. Consequently, the analysis of the time-resolved XRPD patterns is often limited to the quantification of the crystalline fractions and occurrence of an amorphous or poorly crystalline component cannot be easily reckoned and quantified.
In this thesis, the setting reaction was investigated by means of *in-situ* synchrotron X-ray powder diffraction (XRPD). Therefore, in order to study the reaction kinetics and mechanisms, the identification and the description of the time evolution of the crystalline, as well as the amorphous fractions in the sample, were pursued. The information contained in the diffraction pattern, in terms of intensity and Bragg angle 2θ was evaluated through a semi-automated full-profile approach based on multivariate techniques proposed for the analysis of large XRPD datasets, as those obtained from time-resolved *in situ* experiments.

7.1 Experimental

The experiments were carried out at the European Synchrotron Radiation Facility (ESRF), Grenoble (France) with the technical support of the Institute of Theoretical and Applied Mechanics ASCR, Centre of Excellence Telč, (Czech Republic).

7.1.1 Sample preparation

MgO powder obtained by calcination of pharmaceutical grade $MgCO_3$ at 1400 °C was mixed with KDP by hand in agate mortar at unity molar ratio and then placed in a capillary 0.7 mm in diameter opened on both ends. The powder was laterally confined between 2 small layers of quartz wool, allowing for the water to flow through the capillary. The capillary was mounted on a goniometric head for data collection with one end

connected to a vacuum pump. An amount of water to ensure a water/solid weight ratio 1 was introduced on the other end, but initially not in contact with the powders. After starting data collection, operating the vacuum pump, the water was gently allowed to flow through the capillary and wet the powder, defining the start of the experiment. The whole process was followed through a high resolution camera. The advantage of such experimental setup, described in more detail elsewhere (Conterosito *et al.*, 2013), was that the reaction can be monitored from the very beginning. The downsides were that the exact water/solid ratio and the homogeneous wetting of the powder were not assured.

7.1.2 *In-situ* synchrotron powder diffraction

In situ XRPD data were collected at the beamline BM01a, European Synchrotron Radiation Facility (ESRF), Grenoble (France), employing a wavelength of 0.6895 Å with the pilatus 2M detector (Dectris). Isothermal runs at room temperature (20 °C) were conducted allowing the capillary to swing on its axis of 60° following the reaction for 39 min collecting four scans/min. Each spectrum was recorded covering an angular range 1-43.8° 2θ with a resolution of 0.0146° 2θ.

7.1.3 Peak fitting procedures

Transformation conversion curves, expressed as degree of conversion α vs. time, have been built integrating the area of the (200) diffraction peak of MgO, and the (110) of MKP, as described in Viani *et al.* (2016). Single peak integration, accomplished with the software PeakFit (Systat Software Inc.), was dictated by the complex shape of MKP diffraction peaks. For this reason, their shape was described employing 3 pseudoVoigt functions, subtracting a flat background, whereas, 1 pseudoVoigt function, subtracting a flat background, was employed to describe MgO diffraction peaks. KDP, whose diffraction peaks disappeared within 3 minutes from the start of the experiment, was not considered. For MgO decomposition, the integrated area of the MgO peak in the first collected pattern was taken as α=1. Total conversion (α=0) was assumed as the asymptotic value attained by fitting the final part of the conversion curve with an exponential decay function (Viani *et al.*, 2016). In the case of MKP

crystallization, the value of α=0 is easily set as no MKP peaks in the powder pattern are initially observed. The value corresponding to α=1 was set as the asymptotic value of the crystallization curve fitted with an exponential function rising to a maximum. Fitting was accomplished with software SigmaPlot v12 (Systat Software Inc.).

7.1.4 Experimental XRPD patterns

The most significant experimental XRPD patterns were selected by visual inspection and depicted in Figure 45. A comparison with the theoretical patterns of the phases present in the system may suggest the occurrence of the following main steps:

a) before the injection of the water the pattern was characterized only by the reflections pertaining to MgO and KDP (Figure 45 (i));

b) at t ≈ 0 min (water injection), a broad peak due to the diffuse scattering of water and centered at about 11° 2θ (Bergold *et al.*, 2013; Petkov *et al.*, 2005) appeared. As soon as the water was injected, KDP quickly dissolved, and MgO started to react with the solution; the intensity of the background at about 13° 2θ increased (see magnification of Figure 45 (ii) on the top). This was likely the effect of the development of the amorphous product of the reaction;

c) at t ≈ 3.0 min, only magnesia and amorphous contributions can be observed, as the crystalline fraction of the KDP dissolved completely (Figure 45 (iii)). As water was progressively consumed its contribution to the background decreased, conversely, that due to the amorphous product increased, changing the shape of the background;

d) at t ≈ 12.5 min, crystallization of MKP crystals took place (Figure 45 (iv));

e) from t ≈ 27.5 to 39 min (end of the experiment) the crystallization of MKP continued extremely slowly and no other relevant changes were observed (Figure 45 (v)).

The dynamics of the process here conjectured were compared with the results provided by the procedure in order to assess whether they were consistent or not. For our purposes, the data considered during the implementation of the algorithms included patterns of the solid fractions, collected before the water injection for the XRPD patterns in the angular range from 6 to 43.8° 2θ and the time interval from 0 to 39 min.

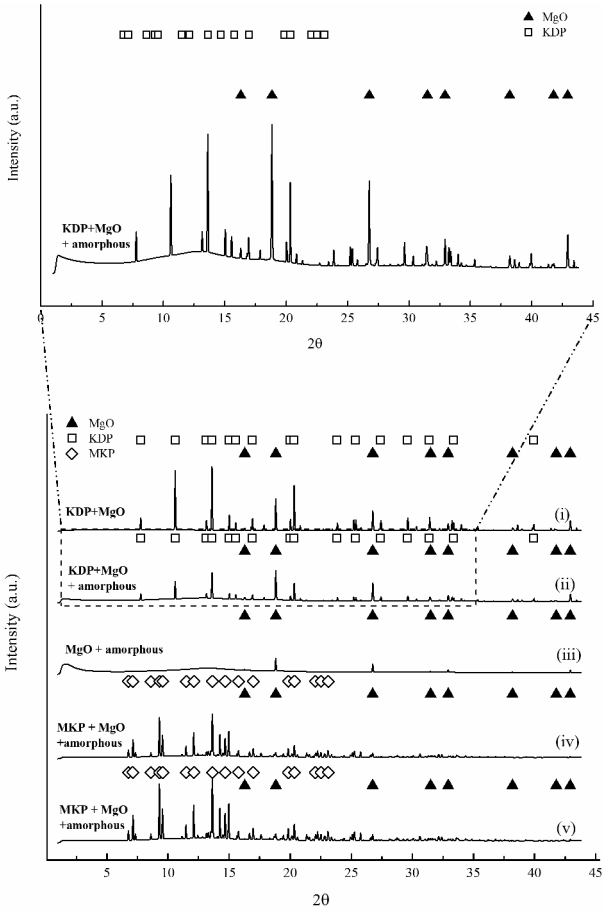

Figure 45 - Patterns representing the most significant changes over the time compared to the theoretical XRPD pattern of MgO, KDP and MKP. Patterns corresponded to: (i) MgO and KDP (before the injection of water), (ii) MgO, KDP and water/amorphous (0 min), (iii) MgO and water/amorphous (3.0 min), (iv) MKP, MgO and water/amorphous (12.5 min) and (v) MKP, MgO and water/amorphous (39 min). Magnification of the pattern (ii) is depicted on the top.

7.2 Methods

7.2.1 Time Window Statistical Total Correlation Spectroscopy

As seen in section 2.10, Statistical Total Correlation Spectroscopy can be suited to any analytical techniques that determine molecular or atomic features. In this thesis, it was employed for the estimation of crystalline phases patterns from X-ray synchrotron powder diffraction data. Up to our knowledge, it has not been employed for XRPD data treatment yet. However, some problems may however arise when driver peaks belong to more than one compound. Particularly, in case of evolving systems, STOCSY cannot accurately estimate compounds patterns when the peaks of reactants and products overlap. Thus, the Time-Window Statistical Total Correlation Spectroscopy (TWSTOCSY) was here introduced for the analysis of patterns collected during in-situ X-ray synchrotron powder diffraction experiment. Particularly, it was supposed to be able to determine the patterns of the crystalline phases, even in case of overlapping peaks. The idea of applying the time window analysis to STOCSY, was based on the works of Manne (1995) and Manne et *al.* (1999) where the sub-windows are employed to determine spectra of the species in hyphenated chromatography. A sub-window is defined as a time interval where a particular species exists and whose amount may vary.

The TWSTOCSY here proposed consisted of the following steps:

1. determine the number K and related size I_K of the time windows. Each time window was associated to a different step of the reaction occurring in the process. This could be carried out by exploiting the EFA results;
2. partition of the data matrix $\mathbf{X}_{(I\times J)}$ in K different submatrices $\mathbf{X}^K_{(IK\times J)}$ where $\sum I^K = I$ and choice of the representative spectrum $\mathbf{x}^K$ for each sub-window;
3. implementation of the STOCSY and evaluation of the A_k patterns pertaining to each submatrix $\mathbf{X}^K$.

4. The procedure was iterated for each window. Estimation of the spectra $\hat{\mathbf{s}}_{p,TWST}$ was eventually carried out (with p=1, …, a_T where $a_T = \sum_{k=1}^{K} A_k$).

TWSTOCSY could allow for detecting the number of species involved in the reaction even in the case of rank-deficient systems. The advantage of using TWSTOCSY over the classical STOCSY was that it could be applied for the resolution of overlapping peaks, since it was able to isolate the interaction between the compounds (or fractions) that exist in different time windows. As a result, the accuracy of the pattern matching could be greatly enhanced.

7.2.2 Multivariate techniques exploited

A semi-automated method combining different multivariate techniques was applied to the analysis of large datasets of time resolved X-Ray powder diffraction patterns. Time-Window Statistical Total Correlation Spectroscopy was here proposed for the pattern matching of the crystalline phase, to be used in case of overlapping peaks. Furthermore, Evolving Factor Analysis and Multivariate Curve Resolution were employed for the identification and the description of the time evolution of the crystalline, as well as the amorphous fractions in the sample.

7.3 Results

7.3.1 Analysis of data with EFA and MCR

EFA was firstly implemented in order to discriminate the main changes occurring in the patterns and determine the raw concentration curves. In the following, the **T** matrix will refer to the conversion of each phase, while the **P** matrix will indicate the diffraction patterns matrix. For the case at hand, according to Equation (27), only two components should in principle be identified (de Juan et *al.*, 2004; Amrhein et *al.*, 1996), since one main reaction should take place. However, according to Figure 46, up to four components could be taken into account, since it showed that four

singular values profiles clearly arose from the noisy profiles depicted in the bottom of the figure. Nevertheless, it was verified that a fourth component did not lead to physically reasonable results. Therefore, three components were believed to capture most of the variance of the matrix **X**.

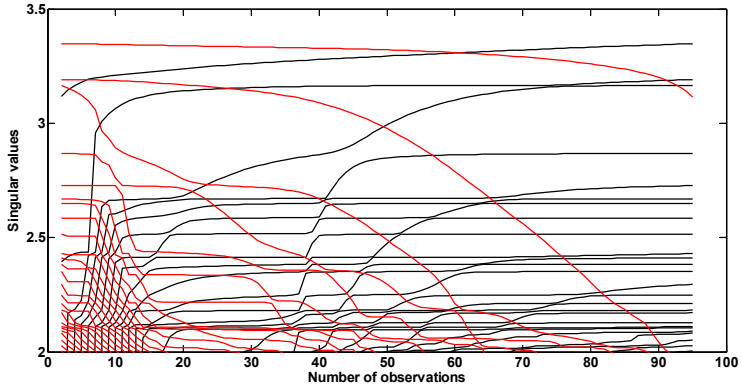

Figure 46 – Singular values estimated as more observations were included in the augmented matrix during EFA algorithm implementation. Black lines refer to the results obtained for the forward EFA, while red lines to the backward EFA.

The time evolution of the components estimated with the EFA procedure is reported in Figure 47a. The factors seemed to behave as: (i) a reactant (open triangles), (ii) intermediate species (black triangles), (iii) a product (grey diamonds). One should note that the behaviour of reactants and products is not monotone. In more detail, both reactants and product show a maximum at t ~ -1.5 min (i.e. before the water injection) and t ~ 38 min, respectively. To estimate both the matrices **T** and **P**, the MCR-ALS algorithm was used by considering the conversion curves previously estimated with EFA as the initial guess. The constraints considered were non-negativity of spectra and concentration matrices and closure of the concentration profiles set to be less than or equal to 1. The outcomes were hereafter referred as the matrices $\widetilde{\mathbf{T}}_{(95\times3)}$ and $\widetilde{\mathbf{P}}_{(1295\times3)}$. As shown in Figure 47b, the MCR-ALS algorithm provided results more plausible than the ones obtained with the EFA. Indeed, first and third components, $\widetilde{\mathbf{t}}_1$ and $\widetilde{\mathbf{t}}_3$, were characterized by a monotonic trend. Furthermore, the second component $\widetilde{\mathbf{t}}_2$ exhibited a sudden increase at *t*~0, then it appeared as

approximately constant for a wide time range (from 7.5 to 17.5 min) before decreasing. The goodness of fit was confirmed by the percentage of variance and the lack of fit (see Equations (32) and (33)) that were equal to 99.8 % and 3.5 %, respectively.

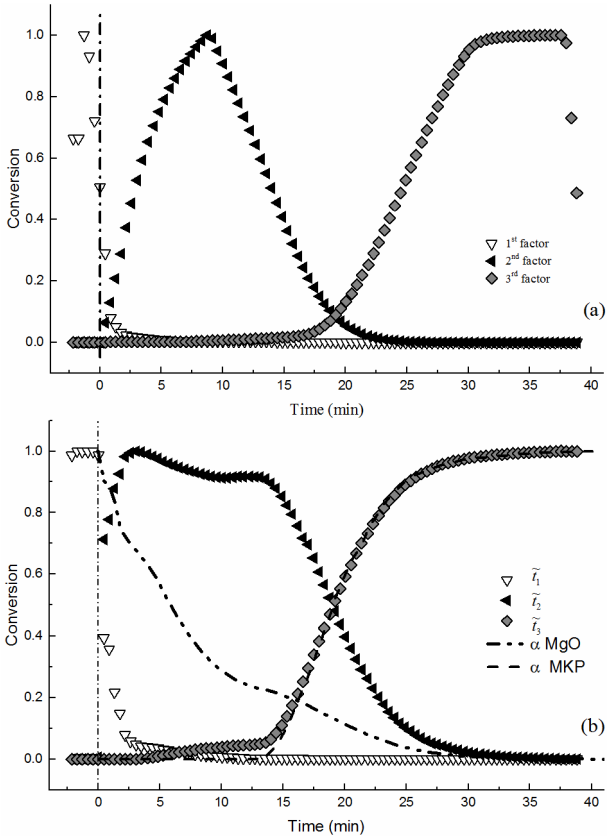

Figure 47 - (a) Normalized conversion curves of the three components estimated with the EFA algorithm. (b) Conversion curves $\widetilde{\mathbf{T}}$ estimated through MCR algorithm, where $\widetilde{\mathbf{t}}_1$ corresponds to reactant, $\widetilde{\mathbf{t}}_2$ to intermediate phase, $\widetilde{\mathbf{t}}_3$ to product conversion curves.

The spectra $\widetilde{\mathbf{P}}$ were compared with the theoretical pattern of MgO, KDP and MKP (Figure 48). It was found that (i) $\widetilde{\mathbf{p}}_1$ corresponded to the sum of MgO and KDP spectra; (ii) $\widetilde{\mathbf{p}}_2$ to MgO and intermediate species (likely the water and amorphous product of reaction); (iii) $\widetilde{\mathbf{p}}_3$ to MKP, in addition a small contribution of the MgO peaks was still appreciated. Hence, the

role of these two algorithms was complementary: on one hand, the EFA helped the identification of compounds and initial guesses, whereas MCR-ALS led to a significant improvement of the conversion curves and spectra estimation. It should be remarked that, for the case at hand, EFA revealed fundamental to detect the presence of the amorphous fractions otherwise not easily determined with the conventional approach.
Nevertheless, as already mentioned, the quality of MCR outcomes depends on the initial estimates $\mathbf{T}^0$ (or, alternatively, $\mathbf{P}^0$). Indeed, for the case under investigation, two reactants spectra (MgO and KDP) collapsed within a common spectrum, since they decreased together and EFA could not distinguish them. Moreover, the presence of unreacted magnesia also precluded the algorithm from separating MgO from MKP and amorphous patterns. Therefore, in order to improve the estimation of spectra of the crystalline fractions with respect to EFA-MCR results, TWSTOCSY algorithm was employed. It aimed to distinguish accurately the different patterns that overlap over the time.

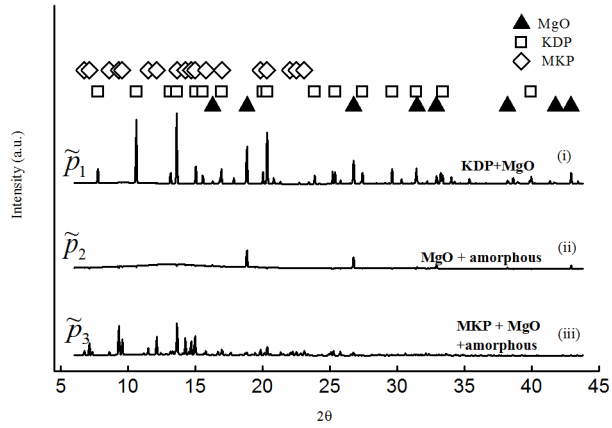

Figure 48 – Spectra $\tilde{\mathbf{P}}$ estimated through MCR algorithm with the theoretical pattern of MgO, KDP and MKP. $\tilde{\mathbf{p}}_1$ corresponds to KDP and MgO, $\tilde{\mathbf{p}}_2$ to amorphous phase and MgO, $\tilde{\mathbf{p}}_3$ to MKP and MgO.

7.3.2 Spectra estimation through TWSTOCSY

Before implementing the TWSTOCSY, patterns were baseline corrected with a linear interpolation function (algorithm developed by Hrovat, 2009) in order to analyse only the contribution coming from the crystalline

fraction. The procedure here proposed differed from the classical STOCSY by taking into account the possible change of correlation with respect to the time.

To this aim, the pattern matching carried out using the TWSTOCSY will be compared with the results obtained through the classical STOCSY. Indeed, during the implementation of the classical STOCSY, the accuracy of the pattern estimation could be affected by occurrence of driver peak in common with more species. The case at hand represents an illustrative example of such scenario, since some peaks belonging to KDP overlapped with the MKP ones (e.g. 2θ= 13.64°, 16.94°, 20.33°). A better insight of this issue is provided in Figure 49, where the evolution of the intensity of three peaks (2θ =7.7°, 10.6° and 13.64°) belonging to KDP, was compared. It can be noted that peaks at 7.7° and 10.6° decreased together (Figure 49a), while a different trend was observed for the peak at 2θ=13.64° that started to increase after 15 minutes, due to the onset of MKP crystal growth. Indeed, this peak was in common between KDP and MKP. The influence of the peak overlap on the correlation coefficient could be quantified by resorting to a local estimation of this statistic. For this purpose, based on concentration profiles shown in Figure 47b, the window size was 6 minutes. It was moved across the data matrix **X** along the time direction (in this case along the columns) and **X** was partitioned in different submatrices $\mathbf{X}^{k}_{(15\times1295)}$ (for k =1, … , 65). The correlation coefficient between peaks at 7.7° and 10.6° 2θ, which was calculated for each submatrix $\mathbf{X}^{k}$, turned out to be positive over the time and greater than 0.95 (open triangles in Figure 49b). On the other hand, it was apparent that the correlation between peaks at 10.6° and 13.64° changed sign with respect to the time (grey circles in Figure 49b). Therefore, the correlation coefficient will be lower than expected if it is computed on the whole dataset.

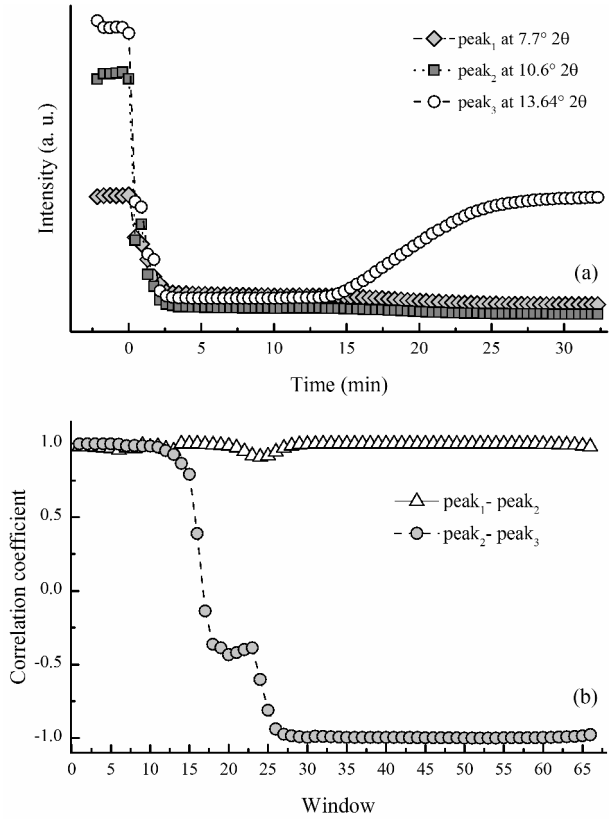

Figure 49 - (a) Evolution of the intensity of three peaks belonging to KDP and (b) correlation coefficient over the time calculated between peaks at 10.6° and 7.7° 2θ (open triangles) and peaks at 10.6° and 13.64° 2θ (grey circles).

Figure 50 compares the correlation coefficient corresponding to the peaks selected on the base of the STOCSY procedure and TWSTOCSY, in case the reactant (KDP) and the product (MKP) have the same driver peak. Indeed, when the peak at 13.64° was considered as the driver one, only few peaks (dashed grey bars at 2θ =15, 16.9, 20.33°) resulted correctly classified as belonging to KDP (Figure 50a), whereas the others were not associated to that (in more detail the peaks at 2θ = 7.76, 10.6, 29.6, 31.4°, reported with grey bars in the figure). Similarly, when the peak at 13.6° 2θ was selected as the driver one, most of the MKP peaks (2θ = 7.12, 9.31, 12.14, 15, 20.33°) were poorly correlated with the driver one (grey bars in Figure 50b) although they were recognized in the literature as belonging

to MKP. As a consequence, the spectrum of the final product was inadequately characterized.

In order to overcome this issue, two sub-windows could be considered: (i) before the water injection until 3.5 min (for a proper estimation of the reactant patterns) and (ii) t=13-39 min (for the products patterns). A qualitative evaluation of the time windows could be suggested by the conversion curves obtained with the EFA and depicted in Figure 47a. The experimental data matrix **X** was then partitioned in two submatrices $\mathbf{X}^1_{(13\times1295)}$ and $\mathbf{X}^2_{(61\times1295)}$. As a result, the correlation with the driver peak at 13.64° 2θ calculated in the first sub-window was higher than the threshold value (η=0.99) and all the peaks were correctly assigned to the KDP pattern, as it can be appreciated from Figure 50a (black bars). As for the KDP pattern estimation, the pattern matching of MKP could be considerably improved calculating the correlation only considering the second sub-window. This eventually led to a good estimation of the MKP patterns as shown in Figure 50b (black bars).

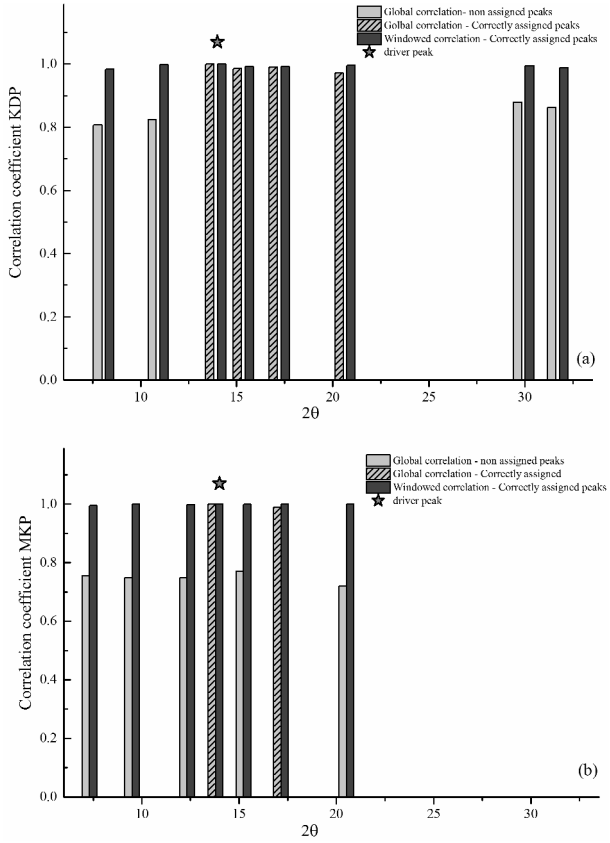

Figure 50 - Correlation coefficient determined considering the whole dataset compared with the one evaluated for one window in case of KDP (a) and MKP (b) patterns had to be identified. Since the driver peak at 13.64° 2θ was common to KDP and MKP, the pattern matching could be improved if the window analysis was carried out (black bars), conversely if the whole dataset was taken into account (grey bars) only few peaks were correctly assigned to KDP or MKP (grey dashed bars).

As seen previously, the TWSTOCSY demonstrated to carry out the pattern matching more accurately than the STOCSY. Therefore, based on the correlation matrix previously determined for each window, $\mathbf{X}^1_{(13\times1295)}$ and $\mathbf{X}^2_{(61\times1295)}$ and proceeding with the steps introduced in Appendix A.1, the full pattern of each phase could be estimated. The results of the procedure for the estimation of the XRPD pattern of the solid reactants are reported in Figure 51. The first spectrum referring to the pure solid fractions is considered as the reference one. When the peak at 2θ = 13.64° was selected as the driver peak (Figure 52(i)), the extracted spectrum

$\hat{\mathbf{p}}_{1,TWST}$ correctly corresponded to the KDP ones (Figure 51(ii)). After removing the KDP contribution, the residual spectrum was the new reference (Figure 51(iii)) and the peak located at 18.84° 2θ (white star) was chosen as the second driver peak. It turned out that the extracted spectrum $\hat{\mathbf{p}}_{2,TWST}$ matched exactly the MgO profile (Figure 51(iv)).

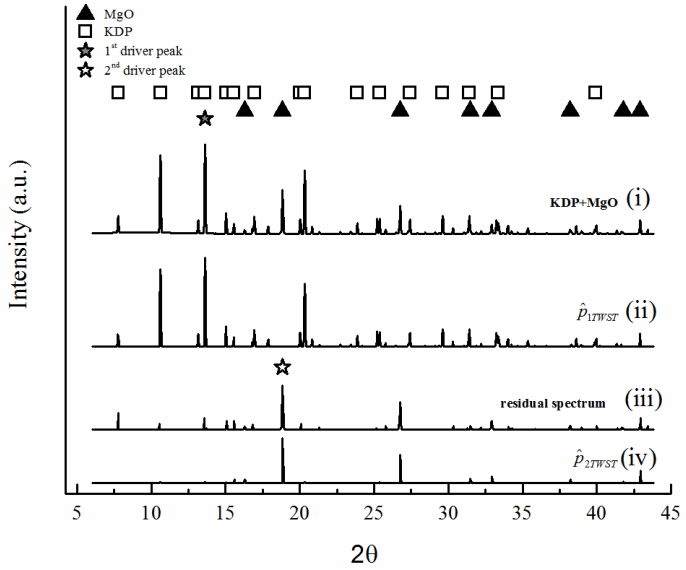

Figure 51 - Spectra extracted through TWSTOCSY applied to the first sub-window. The pattern of solid reactants collected before the water injection is shown (i), in the first iteration the peak at 13.64° 2θ was selected as the driver (grey star); (ii) the $\hat{\mathbf{p}}_{1,TWST}$ pattern was estimated and was removed from the reference spectrum. The first residual spectrum (iii) showed a high intensity peak at 18.84° 2θ (white star) that allowed the extraction of the $\hat{\mathbf{p}}_{2,TWST}$ pattern, reported in Figure 51(iv). For sake of comparison the theoretical patterns of MgO and KDP are also reported.

By analogy, the above procedure was applied to the submatrix $\mathbf{X}^2$ to extract the product pattern. In this case, the pattern recorded at time t=39 min was considered as the reference one (Figure 52(i)). Then, selecting the driver peak at 13.64° 2θ (grey star), the spectrum $\hat{\mathbf{p}}_{3,TWST}$ was determined (Figure 52(ii)), which corresponded to the MKP one. The residual spectrum showed few positive peaks that belonged to unreacted magnesia (Figure 52(iii)).

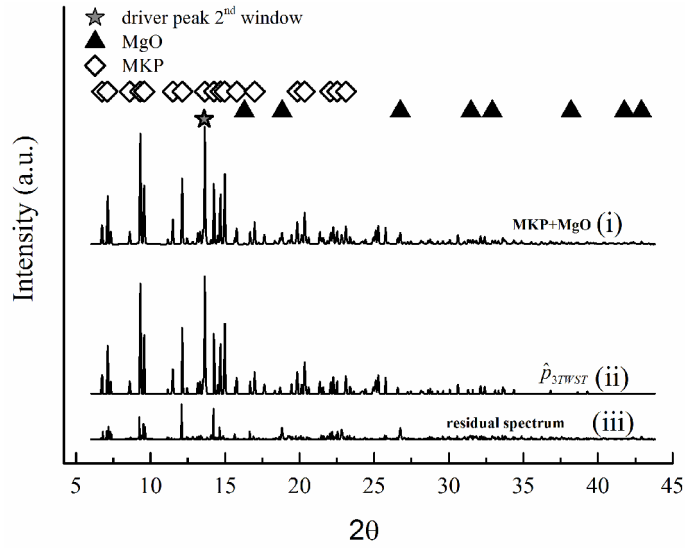

Figure 52 - Spectra extracted through TWSTOCSY applied to the second sub-window. (i) pattern collected at t=39 min; (ii) estimated spectrum $\hat{\mathbf{p}}_{3,TWST}$; (iii) residual spectrum. The theoretical pattern of MgO and MKP is reported to validate the results.

By comparison with theoretical patterns, one can infer that patterns determined with this approach agreed very well with MgO, KDP and MKP. The results were summarized in Figure 53a(i), (ii), (iii), respectively. Indeed, TWSTOCSY allowed for clearly distinguishing magnesia from KDP profile and removing it from MKP one. These results encouraged the application and if needed, the adaptation of the procedure to other dataset where the choice and the partition in different subwindows is not so trivial. However, it is worth to remark that this procedure did require limited a-priori knowledge of the compounds patterns, since the selection of the driver peak was based on the highest intensity without other assumption.

7.3.3 Spectra and conversion estimation through MCR

In order to determine conversion curves, the MCR-ALS was implemented. Then, three crystalline compounds spectra $\hat{\mathbf{p}}_{p,TWST}$ determined through TWSTOCSY (Figure 53a (i), (ii), (iii)) were used as

initial guesses. Nevertheless, since TWSTOCSY was developed to extract spectra of only crystalline compounds, the amorphous spectrum $\tilde{\mathbf{p}}_4^*$ was estimated subtracting the MgO peaks from the $\tilde{\mathbf{p}}_2$ spectrum previously obtained with the EFA-MCR-ALS (Figure 53a (iv)).

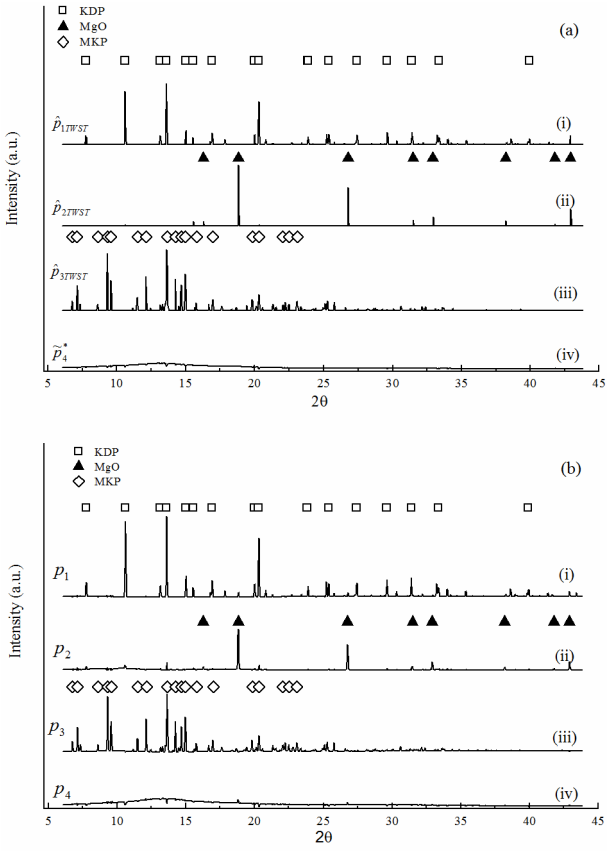

Figure 53 - (a) Patterns used as initial guesses for the MCR-ALS algorithm. Patterns (i), (ii), (iii) were estimated through TWSTOCSY and compared with the theoretical pattern. They corresponded to KDP (i), MgO (ii), and MKP (iii), respectively. Contribution of the incoherent scattering of water and amorphous (iv) obtained through EFA-MCR is reported. (b) Spectra estimated through MCR-ALS that corresponded to (i) KDP, (ii) MgO, (iii) MKP and (iv) water/amorphous with a small contribution of MgO.

In Figure 53b the $\mathbf{p}_a$ spectra (a = 1 to 4) resulting from the MCR-ALS algorithms are reported and compared to the theoretical patterns of the MgO, KDP and MKP. The spectra $\mathbf{p}_1$ to $\mathbf{p}_3$ matched very well the theoretical patterns of KDP (Figure 53b(i)), MgO (Figure 53b(ii)) and

MKP (Figure 53b(iii)), respectively. Moreover, the spectrum $\mathbf{p}_4$ associated to the intermediate phase was also determined (Figure 53b(iv)). Although the estimation of the amorphous phase pattern with the above procedure was rather good, some small positive and negative peaks belonging to the crystalline phases are present in the spectrum. In fact, MCR may encounter problems in perfectly discriminating the different contributions, especially when there are unreacted or inert phases (Amrhein *et al.*, 1996). In this case study, magnesia was not consumed at the end of the experiment. This led MCR to consider the fourth component as a mixture of amorphous and a small amount of MgO. Nevertheless, MCR coupled with TWSTOCSY showed to provide smoother amorphous spectra, at least when compared to the one provided by the simple EFA and MCR protocol. Indeed, the intensity of MgO peak in spectrum $\mathbf{s}_4$ (Figure 53b) was about five times lower than in spectrum $\widetilde{\mathbf{p}}_2$ (Figure 48). As a final remark, one can notice that the algorithm was able to separate the reactants spectra and that could not be pursued by resorting to the EFA-MCR procedure.

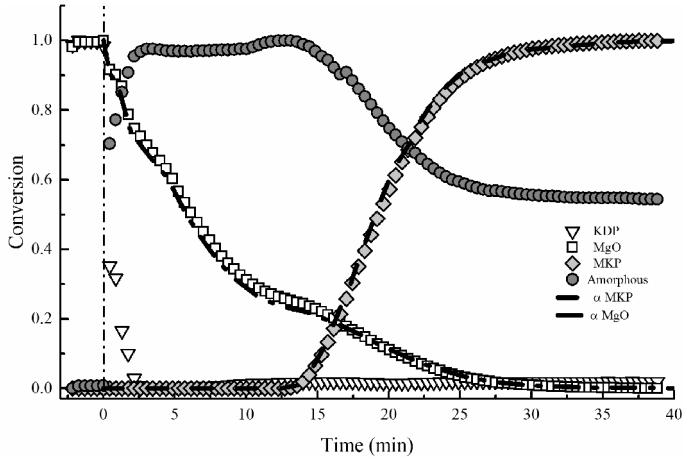

Figure 54 - Conversion curves of MgO, KDP, MKP and the incoherent scattering of water/amorphous phase determined through MCR-ALS. Dashed and dashed-dotted lines represent the estimated α conversion of MKP and MgO, obtained from integrated area of diffraction peaks, respectively.

It is worth noting that the TWSTOCSY allowed a more accurate determination of the conversion curves. The better performance of the TWSTOCSY-MCR with the respect to EFA-MCR was also confirmed by

the decrease of the lack of fit from 3.5 % to 1.8 %, while the percentage of explained variance was almost unity (99.99 %). As a consequence, the description of the reaction dynamics turned out more reliable compared to the outcomes of EFA-MCR discussed in Figure 47. The conversions $\mathbf{t}_a$ (a=1, …, 4) were obtained by normalizing the outcomes of the MCR and they are reported in Figure 54, where five main changes could be identified:

I. Before the water injection, only MgO and KDP crystalline fraction were observed;

II. t=0÷3 *min*: after the water was supplied, the KDP rapidly decreased. This was clearly due to its dissolution. The contribution of the incoherent scattering from water summed to the amorphous phase and quickly increased, reaching a maximum at t ~ 3 *min*;

III. t=3÷13 *min*: MgO continued dissolving. The change of slope observed at t ~ 10 *min* could be explained with the reactivation of the reaction in some areas of the sample. This may be likely due to a non-homogeneous wetting of the powder during water injection. On the other hand, the amorphous phase curve appeared to be constant for this time interval. As more amorphous formed, water was consumed, and both signals were summed up. Meanwhile the gel phase required time to thicken and then to nucleate and crystallize in the product (Wagh & Jeong, 2003). This is the well-known induction period in case of crystallization from amorphous (Bhugra & Pikal, 2008; Šimon *et al.*, 2005).

IV. t=13÷27 min: MKP crystallization took place. Simultaneously, the amorphous content decreased;

V. t > 27 min: crystal growth rate significantly slowed down and the amorphous decay flattened out. Thus, a relevant amount of amorphous fraction was still present at 39 min. This behaviour was in agreement with previous results (Viani & Gualtieri, 2014; Viani *et al.*, 2015; Viani *et al.*, 2016).

One can note that the conversion curves depicted in Figure 54 confirmed the interpretation of the reaction mechanism proposed in section 7.1.4, furthermore one of advantages of the approach introduced here consisted in being less time consuming than visually assessing all the patterns collected over the time and identifying the different phases present in each pattern. For sake of comparison, the conversion curves of magnesia and MKP obtained from the integration of single diffraction peaks described in Section 7.1.3 were depicted in Figure 54 as dashed and dashed-dotted lines, respectively. It can be seen that the curves evaluated through TWSTOCSY-MCR agreed very well with them. Particularly, the consistency of this result was also confirmed by the Pearson's correlation between the conversion curves and the estimated ones (see Equation (35)): it was equal to 0.9995 and 0.9998 for MgO and MKP, respectively. It should be also pointed out that the proposed procedure was able to identify the non-crystalline component (depicted as circles in Figure 54) not accounted for by the traditional approach. In the first minutes, this component included the contribution of the liquid solution. By the time MKP crystallization started, and MgO dissolution was close to completion, the signal could be considered to come largely from the solid amorphous precursor, already observed, even at longer times, by SEM and detected with QPA with the external standard method (Viani & Gualtieri, 2014; Viani *et al.*, 2015). The synchronous onset of MKP crystallization and decrease of the non-crystalline component, processes occurring with the same dynamic, supported the view that MKP crystallizes from an amorphous precursor. In this work no separation of the contribution of the incoherent scattering of water from that of the amorphous precursor, has been attempted. As a final remark, the procedure introduced in this work not only allowed for the pattern matching and the evaluation of the conversion curves with limited or no user intervention, but it also resulted less time consuming than the conventional available methods.

In conclusion, the main features of the procedure proposed to investigate the setting reaction of MKP and treat X-Ray powder diffraction data were summarized below: (i) it was a full-profile approach i.e. it took into account the whole powder diffraction profile rather than focusing on a limited number of peaks; (ii) it required a limited knowledge of the spectra of the pure phases; (iii) the windowed analysis allowed for

isolating also phases that existed in different time windows but had overlapping peaks (such as reactants and products). In the case under investigation, TWSTOCSY demonstrated to be an excellent tool to distinguish the pattern of KDP from MKP and MgO ones. In addition, when combined with other techniques (i.e. the Multivariate Curve Resolution and the Evolving Factor Analysis), it was able to describe the evolution of the non-crystalline component, not easily appreciable using other methods for the analysis of XRPD data. Conversion curves were accurately reproduced, allowing for the study of reaction kinetics and identification of reaction mechanisms. These results encouraged to validate the procedure considering other evolving system where the windowing analysis could result not trivial. Although it was applied to investigate the setting reaction of a ceramic material, it can be extended to data also coming from other analytical spectroscopic techniques.

Chapter 8

Conclusions

The main conclusions of the present work are summarized in this Chapter.

In this thesis procedures based on multivariate techniques were proposed for on-line monitoring of continuous and evolving processes.
With this regard, spectroscopic measurements demonstrated an extremely useful tool to extract chemical information from data collected.
The on-line monitoring of continuous processes such as commercial detergents production was pursued through procedures based on Principal Component Analysis and Partial Least Squares Regression. In more detail, the Elliptical Normal Operating Region (ENOR) was developed by means of a Box-Cox transformation of the Q statistic, to detect occurrence of deviations from the reference state. The ENOR demonstrated to be capable of discriminating the in-control from the out-of-control observations with a high sensitivity. In addition, concentrations of the surfactant were estimated through Partial Least Squares Regression, with a quite high prediction performance. The reliability of the estimation was evaluated through the Q_x statistic: when samples were classified as out-of-control, the model could not be employed for the estimation.
Concerning the monitoring of evolving processes, Moving Window Principal Component Analysis was employed. It is worth noting that this method has been seldom applied to *in situ* spectroscopic data. The crystallization of Isonicotinamide examined by means of infrared spectroscopy was considered as a representative case study. Therefore, the nucleation temperature was correctly detected through T^2 and Q control chart (T= -7.86 °C). The proposed approach noticeably reduced the false positive ratio to 5.8 %, at least when compared to the 77 % of false positives observed with the static PCA). Eventually, the contribution plot could identify which were the spectral ranges mostly affected by nucleation.
Phenomena occurring in evolving systems could be investigated through multivariate techniques as well. For this purpose, two case studies were investigated: (i) dissolution of surfactants paste studied by means of confocal Raman microscopy and (ii) the setting reaction of magnesium potassium phosphate (MKP) analyzed through X–ray powder diffraction.
For the former case, Multivariate Curve Resolution was suggested for data treatment. The estimated spectra well reproduced the experimental ones. MCR concentration profiles of paste A and B were fitted through an exponential decay model and the estimated dissolution rate could suggest that both surfactants were dissolving at the same rate.

Concerning the setting reaction, Time Window Statistical Total Correlation Spectroscopy combined with Multivariate Curve Resolution (MCR) was here proposed. The main advantages over the conventional approaches were the following: the procedure required a limited knowledge of the spectra of the pure phases and particularly TWSTOCSY allowed for isolating also crystalline phases that existed in different time windows but had overlapping peaks (such as reactants and products). The combined procedure showed to distinguish the four components involved (MgO, KDP, MKP and amorphous phase) not otherwise possible through the only Evolving Factor Analysis-MCR. The estimation of their spectra and concentration was quite good, indeed the time evolution of MgO and MKP agreed very well with the conventional conversion curves.

In summary, the potentialities of multivariate methods were illustrated in this thesis. The variety of system examined demonstrated the flexibility and the usefulness of the multivariate techniques employed.
As it has been noted, both spectroscopy and multivariate techniques can be used for process analysis and monitoring, therefore they could lead to improve control system reliability and process performances. As a result, product quality can be monitored online and energy and costs could be saved. On the other hand, they showed as essential tools to investigate system dynamics and phenomena, not otherwise possible with conventional experimental and mathematical techniques.

References

Akaike H. (1974). IEEE Transactions on Automatic Control. AC-19, 716–723.

Alcalà M., Blanco M., Bautista M. & González J. M. (2010). Journal of Pharmaceutical Sciences, 99 (1), 336-345.

Alcala C. F. & Qin S. J. (2011). Journal of Process Control. 21, 322–330.

Alexandrino G. L., Khorasani M. R., Amigo J. M., Rantanen J. & Poppi R. J. (2015). European Journal of Pharmaceutics and Biopharmaceutics 93, 224–230.

Allen S. & Evans J. S. (2004). Journal of Material Chemistry. 14, 151-156.

Allen S., Warmingham N. R., Gover R. K. & Evans J. S. (2003). Chemistry of Materials. 15, 3406-3410.

Amrhein M., Srinivasan B., Bonvin D. & Schumacher M. M. (1996). Chemometrics and Intelligent Laboratory Systems. 33, 17-33.

Anderson E., Bai Z., Bischof C., Blackford S., Demmel J., Dongarra J., Du Croz J., Greenbaum A., Hammarling S., McKenney A. & Sorensen D. (1999). LAPACK User's Guide, SIAM, Philadelphia.

Armstrong R. A. (2014). Ophthalmic & Physiological Optics 34, 502–508.

Bates S., Zografi G., Engers D., Morris K., Crowley K. & Newman A. (2006). Pharmaceutical Research. 23, 2333-2349.

Bergold S. T., Goetz-Neunhoeffer F. & Neubauer J. (2013). Cement and Concrete Research. 53, 119-126.

Bhugra C. & Pikal M. J. (2008). Journal of Pharmaceutical Sciences. 97, 1329-1349.

Bish D. L. & Howard S. A. (1988). Journal of Applied Crystallography. 21, 86-91.

Bolze J., Peng B., Dingenouts N., Panine P., Narayanan T. & Ballauff M. (2002). Langmuir. 18, 8364-8369.

Borin A. & Poppi R.J. (2007). Journal of Brazilian Chemical Society. 15(4), 570-576.

Brereton R.G. (2000). Analyst. 125, 2125–2154.

Broadhurst D.I. & Kell D.B. (2006). Metabolomics. 2(4), 171–196.

Brown M.E., Dollimore D. & Galway A.K. (1980). Theory of Solid State Reaction Kinetics, Compr. Chem. Kinet. 22, React. Solid State, 41–113.

Bu D., Wan B. & McGeorge G. (2013). Chemometrics and Intelligent Laboratory Systems 120. 84–91.

Burud I., Gobakken L. R., Flø A., Kvaal K. & Thiis T. K. (2014). International Biodeterioration & Biodegradation 88, 37-43.

Camacho J., Picó J. & Ferrer A. (2008). Journal of Chemometrics. 22(5), 299–308.

Cartwright J. H., Checa A. G., Gale J. D., Gebauer D. & Sainz-Díaz C. I. (2012). Angewandte Chemie International Edition. 51, 11960-11970.

Cattaneo A., Gualtieri A. F. & Artioli G. (2003). Physics and Chemistry of Minerals. 30, 177-183.

Chen B-H., Miller C. A. & Garrett P. R. (2001). Colloids and Surfaces A: Physicochemical and Engineering Aspects. 183–185, 191–202.

Chen Q., Kruger U., Meronk M. & Leung A. Y. T. (2004). Control Engineering Practice. 12, 745-755.

Chen Z.-P., Morris J., Borissova A., Khan S., Mahmud T., Penchev R. & Roberts K. J. (2009). Chemometrics and Intelligent Laboratory Systems. 96, 49–58.

Cheng W., Sun D-W & Cheng J-H. (2016). Food Science and Technology. 73, 13-19.

Cloarec O., Dumas M. E., Craig A., Barton R., Trygg J., Hudson J., Blancher C., Gauguier D., Lindon J., Holmes E. & Nicholson J. (2005). Analytical Chemistry. 77, 1282−1289.

Conterosito E., Van Beek W., Palin L., Croce G., Perioli L. & Viterbo, D. (2013). Crystal Growth and Design. 13, 1162-1169.

Costa P. & Lobo J. M. S. (2001). European Journal of Pharmaceutical Sciences. 13, 123–133.

Cozzolino D., Liu L., Cynkar W.U., Dambergs R.G., Janik L., Colby C. B. & Gishen M. (2007). Analytica Chimica Acta. 588, 224–230.

de Jong S. (1993). Chemometrics and intelligent laboratory systems. 18 (3), 251-253.

de Juan, A., Jaumot, J. & Tauler, R. (2014). Analytical Methods, 6, 4964-4976.

de Juan A., Navea S., Diewok J. & Tauler R. (2004). Chemometrics and Intelligent Laboratory Systems. 70, 11–21.

de Juan A., Rutan S. C. & Tauler R. (2009). Comprehensive Chemometrics. Edited by Brown S. D., Tauler R., Walczak B. 2, 325-344. Oxford: Elsevier.

De Ketelaere B., Hubert M. & Schmitt E. (2015). Journal of Quality Technology. 47, 4, 318- 335.

Dorozhkin, S. V. (2010). Acta Biomaterialia. 6, 4457-4475.

Everall N. J. (2009). Applied Spectroscopy. 63 (9), 245A-262A.

Filho O. T., Pinheiro J. C., da Costa E. B., Kondo R. T., de Souza R. A., Nogueira V. M. & Mauro A. E. (2006). Journal of Molecular Structure: THEOCHEM, 763, 175–179.

Garrido M., Rius F. X. & Larrechi M. S. (2008). Analytical and Bioanalytical Chemistry. 390, 2059–2066

Godoy J. L., Vega J. R., Marchetti J. L. (2014). Chemometrics and Intelligent Laboratory Systems 130, 182–191.

Gradzielski M. (2003). Current Opinion in Colloid and Interface Science. 8, 337–345.

Gualtieri A. F., Giacobbe C. & Viti C. (2012). American Mineralogist. 97, 666-680.

Guidance for Industry, PAT—a framework for innovative pharmaceutical manufacturing and quality assurance, U.S. Food and Drug Administration (FDA), Rockville MD, USA, 2004.

Gurden S. P., Westerhuis J. A. & Smilde A. K. (2002). AIChE Journal. 48(10), 2283-2297.

Hall D. A., Stevens R. & El-Jazairi B. (1988). Journal of American Ceramic Society. 81, 1550–1556.

Hantao L. W., Aleme H. G., Pedroso M. P., Sabin G. P., Poppi R. J. & Augusto F. (2012). Analytica Chimica Acta. 731, 11–23.

He X. B. & Yang Y. P. (2008). Industrial & Engineering Chemistry Research. 47, 419– 427.

Hotelling H. (1947). Multivariate quality control, illustrated by the air testing of sample bombsights in Techniques of Statistical Analysis. C. Eisenhart, M.W. Hastay and W.A. Wallis (Editors). 113-184. McGraw-Hill: New York.

Hrovat M. (2009). Baseline fit. Available at: <https://nl.mathworks.com/matlabcentral/fileexchange/24916-baseline-fit> (accessed 20/09/2015).

Hunter J. S. (1986). Journal Quality Technology, 18, 203-210.

Jackson J. E. (1991). A Users Guide to Principal Components, John Wiley & Sons, Inc.: NY, USA.

Jackson J. E. & Muldholkar G. S. Technometrics. (1979). 21, 341-349.

Jansen D., Bergold S. T., Goetz-Neunhoeffer F. & Neubauer J. (2011). Journal of Applied Crystallography. 44, 895-901.

Jaumot J., Gargallo R., de Juan A. & Tauler R. (2005). Chemometrics and Intelligent Laboratory Systems. 76, 101-110.

Jeng J.-C. (2010). Journal of the Taiwan Institute of Chemical Engineer. 44, 475–481.

Joliffe, I. T. (2002). Principal component analysis. New York: Springer.

Kogermann K., Veski P., Rantanen J. & Naelapää K. (2011). European Journal of Pharmaceutical Science. 43, 278–289.

Kourti T. (2006). Analytical and Bioanalytical Chemistry. 384, 1043–1048.

Kourti T. (2005). International Journal of Adaptive Control and Signal Processing. 19, (4), 213-246.

Kruger U., Zhou Y. & Irwin G. (2004). Journal of Process Control 14(8), 879–888.

Ku W., Storer R. H. & Georgakis C. (1995). Chemometrics and Intelligent Laboratory Systems. 30, 179-196.

Kubo T., Ohtani E. & Funakoshi K. I. (2004). American Mineralogist. 89, 285-293.

Kurdowski W. (2014). Cement and Concrete Chemistry. Holland: Springer.

Lee C. A., Gasster S. D., Plaza A., Chang C. I. & Huang B. (2011). International Journal of Applied Earth Observation and Geoinformation. 4, 508-527.

Leineweber A. & Mittemeijer E. (2012). Modern Diffraction Methods. Mittemeijer E. J. & Welzel U. Eds. Weinheim: Wiley-VCH Verlag & Co. KGaA.

Li G., Qin S. J. & Zhoua D. (2010). Automatica, 46(1), 204-210.

Li W., Yue H. H., Valle-Cervantes S. & Qin S. J. (2000). Journal of Process Control. 10, 471-486.

Lilliefors H. W. (1967). Journal of the American Statistical Association. 62, 399–402.

Lin Z., Zhou L., Mahajan A., Song S., Wang T., Ge Z. & Ellison D. (2006). Journal of Pharmaceutical and Biomedical Analysis. 41, 99–104.

Lu N., Yao Y., Gao F. & Wang F. (2005). AIChE Journal. 51(12), 3300-3304.

Lutterotti L., Scardi P. & Maistrelli P. (1992). Journal of Applied Crystallography. 25, 459-462.

MacGregor J. F. & Kourti T. (1995). Control Engineering Practice. 3(3), 403-414.

MacGregor J. F., Jaeckle C., Kiparissides C. & Koutoudi M. (1994). AIChE Journal, 40(5), 826-838.

Maeder M. (1987). Analytical Chemistry. 59, 527–530.

Mandawala A. A., Harvey S. C., Roy T. K. & Fowler K. E. (2016). Animal Reproduction Science. 174, 2–10.

Manne R. (1995). Chemometrics and Intelligent Laboratory Systems. 27, 89-94.

Manne R., Shen H. & Liang Y. (1999). Chemometrics and Intelligent Laboratory Systems. 45, 171-176.

Mazet V., Carteret C., Brie D., Idier J. & Humbert B. (2005). Chemometrics and Intelligent Laboratory Systems. 76, 121– 133.

Meng X., Pan Q., Ding Y & Jiang L. (2014). Food Chemistry. 147, 272–278.

Montgomery D. (2008)a. Design and analysis of experiments. Wiley, Hoboken, NJ, USA.

Montgomery D. (2008)b. Introduction to Statistical Quality Control, Wiley Desktop Editions Series. Wiley: New Jersey.

Mukherjee S., Gowen A. (2015). Analytica Chimica Acta. 895,12-34.

Müller M., Dinnebier R. E., Jansen M., Wiedemann S. & Plüg C. (2009). Powder Diffraction. 24, 191-199.

Nagai S., Ichie T., Yoneyama A., Kobayashi H., Inoue T., Ishii R., Suzuki R. & Itioka T. (2016). Ecological Informatics. 32, 91–106.

Noda I. (1993). Applied Spectroscopy. 47, 1329−1336.

Nomikos P. & MacGregor J. F. (1995). Technometrics, 37, 41–59.

Norby P. & Schwarz U. (2008). Powder diffraction: theory and practice. Dinnebier R. E. & Billinge S. J. Eds. Cambridge: Royal Society of Chemistry.

Nývlt J. (1985). The Kinetics of Industrial Crystallization. Elsevier, New York.

O'Connor B. H. & Raven M. D. (1988). Powder Diffraction. 3, 2-6.

Petkov V., Peng Y., Williams G., Huang B., Tomalia D. & Ren Y. (2005). Physical Review B. 72, 195402.

Pierna A. F., Wahl F., de Noord O. E. & Massart D. L. (2002). Chemometrics and Intelligent Laboratory Systems. 63, 27– 39.

Politi Y., Arad T., Klein E., Weiner S. & Addadi L. (2004). Science. 306, 1161-1164.

Pöllänen K., Häkkinen A., Reinikainen S-P., Rantanen J., Minkkinen P. (2006). Chemometrics and Intelligent Laboratory Systems. 84, 126–133.

Prats-Montalbán J.M., de Juan A. & Ferrer A. (2011). Chemometrics and Intelligent Laboratory Systems. 107, 1–23.

Qin J. (2003). Journal of Chemometrics. 17, 480–502.

Richardson I. G. (1999). Cement and Concrete Research. 29, 1131-1147.

Robinette S. L., Lindon J. C. & Nicholson J. K. (2013). Analytical Chemistry. 85, 5297−5303.

Romagnoli J. & Palazoglou A. (2012). Introduction to process control. Boca Raton, CRC Press.

Sâsić S., Muszynski A. & Ozaki Y. (2000). Journal Physical Chemistry A. 104, 6380-6387.

Savitzky A. & Golay M. J. E. (1964). Analytical Chemistry. 36, 1627–1639.

Scarlett N. V. & Madsen I. C. (2006). Powder Diffraction. 21, 278-284.

Scarlett N. V. Y., Rowles M. R., Wallwork K. S. & Madsen I. C. (2010). Journal of Applied Crystallography. 44, 60-64.

Schaefer C., Lecomte C., Clicq D., Merschaert A., Norrant E. & Fotiadu F. (2013). Journal of Pharmaceutical and Biomedical Analysis. 83, 194 – 201.

Scheipers U., Perrey C., Siebers S., Hansen C. & Ermert H. (2005). Ultrasonic Imaging. 27(3), 181-198.

Schmitt R., Rato T., Ketelaere B., Reis M. & Huberta M. (2016). Journal of Chemometrics. 30, 163–176.

Sciutto G., Oliveri P., Prati S., Quaranta M., Bersani S. & Mazzeo R. (2012). Analytica Chimica Acta. 752, 30–38.

Shewhart W.A. Economic Control of Quality of Manufactured Product, Van Nostrand, Princeton, NJ, 1931.

Simoglou A., Georgieva P., Martin E. B., Morris A. J. & Feyo de Azevedo S. (2005). Computers and Chemical Engineering. 29, 1411–1422.

Šimon P., Nemčeková K., Jóna E., Plško A. & Ondrušová D. (2005). Thermochimica Acta. 428, 11-14.

Simone E., Saleemi A. N. & Nagy Z. K. (2014). Chemical Engineering Research and Design. 92, 594–611.

Solberg C. & Hansen S. (2001). Cement and Concrete Research. 31, 641-646.

Soudée E. & Péra J. (2000). Cement and Concrete Research. 30, 315-321.

Steinke R., Newcomer P., Komarneni S. & Roy R. (1988). Material Research Bulletin. 23, 13-22.

Stinton G. W. & Evans J. S. (2007). Journal of Applied Crystallography. 40, 87-95.

Stuart B. (2004). Infrared spectroscopy: fundamentals and applications. Wiley, Chichester, UK: Wiley.

Szymańska E., Saccenti E., Smilde A. K. & Westerhuis J. A. (2012). Metabolomics. 8, S3–S16.

Tauler R. (1995). Chemometrics and Intelligent Laboratory Systems. 30, 133–146.

Terra L.A. & Poppi R.J. (2014). Chemometrics and Intelligent Laboratory Systems. 130 91–97.

Thirunahari S., Chow P. S. & Tan B. H. R. (2011) Crystal Growth and Design. 11, 3027–3038.

Tracy N. D., Young J. C. & Mason R. L. J. (1992). Quality Technology. 24, 88–95.

van der Meer F.D., van der Werff H. M. A., van Ruitenbeek F. J. A., Hecker C. A., Bakker W. H., Noomen M. F., van der Meijde M., Carranza E. J. M., de Smeth J. B. & Woldai T. (2012). International Journal of Applied Earth Observation and Geoinformation. 14, 112-128.

Viani A. & Gualtieri A. F. (2014). Cement and Concrete. Research. 58, 56–66.

Viani A., Peréz-Estébanez M., Pollastri S. & Gualtieri A. F. (2016). Cement and Concrete Research. 79, 344-352.

Viani A., Radulescu A. & Peréz-Estébanez M. (2015). Materials Letters. 161, 628-630.

Wagh A. S. & Jeong S. Y. (2003). Journal of American Ceramic Society. 86, 1838–1844.

Wagh A. S. (2004). Chemically bonded phosphate ceramics: 21st century materials with diverse applications. Edit by E. Hurst, Amsterdam: Elsevier.

Wang X., Kruger U. & Irwin G.W. (2005). Industrial & Engineering Chemistry Research. 44, 5691–5702.

Warren P. B. & Buchanan M. (2001). Current Opinion in Colloid & Interface Science. 6, 287-293.

Westerhuis J. A., Gurden S. P., Smilde A. K. (2000). Chemometrics and Intelligent Laboratory Systems. 51, 95–114.

Windig W. (1997). Chemometrics and Intelligent Laboratory Systems. 36, 3–16.

Wise B. M. & Gallagher N. B. (1996). Journal of Process Control. 6, 329–348.

Wold H. (1975). Path Models with latent variables: the NIPALS approach, in Quantitative sociology: international perspectives on mathematical and statistical modeling. H. M. Blalock et al. (Eds), Academic, 307-357.

Wold S., Sjöström M. & Eriksson L. (2001). Chemometrics and Intelligent Laboratory Systems. 58, 109–130.

Wold S. (1995). Chemometrics and Intelligent Laboratory Systems. 30(1), 109–115.

Wolf U., Leiberich R. & Seeba J. (1999). Catalysis Today. 49, 411-418.

Wong C. W. L., Escott R., Martin E. & Morris J. (2008). The Canadian Journal of Chemical Engineering. 86, 905–923.

Woodward R. H. & Goldsmith P. L. Cumulative Sum Techniques, Oliver and Boyd, London, 1964.

Yurdakul S., Atac A., Ahin, E. S & De S. I. (2003). Vibrational Spectroscopy. 31, 41-49.

Zeaiter M. & Rutledge D. (2009). Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, from Comprehensive Chemometrics. Edited by Brown S. D., Tauler R. & Walczak B. 3, 121-231. Oxford: Elsevier.

Zhaomin L., Qingchao J. & Xuefeng Y. (2014). Industrial & Engineering Chemistry Research. 53, 6457−6466.

Acknowledgments

I wish to express my sincere gratitude to my supervisor, Massimiliano Grosso, that has provided scientific and personal support to me during these years. It has been extremely constructive and also a pleasure to work under his supervision.
This PhD thesis is the outcome of different collaborations.
I would like to thank Vincenzo Guida, Mariarosa Brundu and Fabio Zonfrilli that work in Procter & Gamble as researchers and proposed very interesting case studies to analyze such as the commercial detergent production and the dish paste dissolution. They supervised me and showed me how the industrial research works during my stay at the Brussel Innovation Center of Procter & Gamble. Thanks to Jurgen Vanpoppel for having carried out the experiments of the dish paste dissolution.
I am also thankful to Alberto Viani, the head of the Laboratory for physical and chemical analyses and material innovations at Centrum Excelence Telč, for his technical support and for having carried out the experiment concerning the setting reaction.
I wish to thank professors Haiyan Qu and Ben-Guang Rong for having hosted me at University of Southern Denmark and for their scientific support. Thanks to Thomas Hansen and the undergraduates students Nete Sloth Bækgaard and Katrine Englund Christensen for all of their experimental work regarding the crystallization of Isonicotinamide, particularly Thomas for the collaboration.

Thanks to my friends and my flatmates for staying with me since long time, to Stefania for the nice conversations.
During these three years, I have met many people and the list would be very long, but I would like to thank my colleagues at the Department of Mechanical, Chemical Engineering and Materials, friends met during my stay at University of Southern Denmark and in Brussels Innovation Center.
Thank you all for your friendliness, kindness and love, for the great moments spent together, I have grown up a lot with you.
Last, I am grateful to my family for their support and love.