UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Industrial Engineering

Ph.D. COURSE IN INDUSTRIAL ENGINEERING

CURRICULUM: CHEMICAL AND ENVIRONMENTAL ENGINEERING

36th SERIES

EFFECTIVE IMPLEMENTATION OF INDUSTRY 4.0 APPROACHES FOR PROCESS MONITORING IN THE BATCH MANUFACTURING OF SPECIALTY CHEMICALS

Coordinator: Ch.mo Prof. Giulio Rosati

Supervisor: Ch.mo Prof. Massimiliano Barolo

Ph.D. student: Francesco Sartori

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Industrial Engineering)
Curriculum: Chemical and Environmental Engineering

at the
University of Padova
2023

A chi mi è stato vicino

Foreword

The fulfilment of the research included in this Dissertation involved the intellectual and financial support of many people and institutions, to whom the author is very grateful.
Most of the research that led to the results reported in this Dissertation has been carried out at CAPE-Lab – Computer-Aided Process Engineering Laboratory, at the Department of Industrial Engineering of the University of Padova (Italy) under the supervision of Prof. Massimiliano Barolo, with significant inputs also from Prof. Fabrizio Bezzo and Prof. Pierantonio Facco. Part of the work was carried out at the BASF Italia SpA site in Pontecchio Marconi (Bologna, Italy) under the supervision of Dr. Matteo Cicciotti and in collaboration with Federico Zuecco; furthermore, part was developed during a 3-month stay at BASF Lampertheim GmbH (Lampertheim, Hessen, Germany), under the supervision of Stefan Höser.
The realization of this work has been made possible through the financial support of BASF Italia S.p.A. (Pontecchio Marconi, Bologna, Italy).

All the material reported in this Dissertation is original, unless explicit references to studies carried out by other people are indicated. In the following, a list of publications stemmed from this project is reported.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS

Sartori F., Zuecco F., Facco, P., Bezzo F., Barolo M. (2022), Data Analytics Can Help Reduce Energy Consumption in the Industrial Manufacturing of Specialty Chemicals. *Chem. Eng. Trans.* **96**, 229-234

Sartori F., Facco, P., Zuecco F., Bezzo F., Barolo M. (2023), Optimal indicator-variable approach for trajectory synchronization in uneven-length multiphase batch processes. *Ind. Eng. Chem. Res.* **62**, 18511-18525

CONTRIBUTIONS SUBMITTED TO INTERNATIONAL JOURNALS

CONTRIBUTIONS IN INTERNATIONAL JOURNALS (in preparation)

Sartori F., Facco, P., Bezzo F., Barolo M. (2023), On the application of assumption-free modelling for multivariate statistical batch process monitoring. *In preparation.*

CONFERENCE PRESENTATIONS

Sartori F., Zuecco F., Facco P., Bezzo F., Barolo M. (2022), Coupling machine learning and engineering judgment to reduce the cycle time of an industrial batch process, Oral presentation at: *GRICU 2022: Centralità dell'Ingegneria Chimica in un Mondo che cambia, 03-06 July, 2022*, (Ischia, NA, Italy)

Sartori F., Zuecco F., Facco P., Bezzo F., Barolo M. (2022), Data Analytics Can Help Reduce Energy Consumption in the Industrial Manufacturing of Specialty Chemicals. Oral presentation at: *1st International Conference on Energy, Environment and Digital Transition (E2DT), October 23-26 2022* (Milano, Italy)

Sartori F., Zuecco F., Facco P., Bezzo F., Barolo M. (2023), An automated pre-processing framework for uneven-length multiphase batch processes. Oral presentation at: *11th Colloquium Chemometricum Mediterraneum, June 27-30 2023* (Padova, Italy)

Abstract

Batch processes are widely used in industrial sectors that produce high-value products due to their ease of setup and operational flexibility. Although they are effective for the manufacturing of relatively small amounts of high value-added products, controlling them to maintain consistently high product quality is more challenging than in continuous processing. Additionally, quality control is usually performed only at the end of a batch, which further complicates quality assurance. The advancements brought about by Industry 4.0 have enabled the monitoring of numerous process variables. However, observing just one variable at a time can overwhelm the process supervisor with information. Multivariate statistical methodologies can alleviate this issue by reducing the complexity of the problem while handling noise, multicollinearity, and missing data. The aim of this PhD project is to develop multivariate statistical methodologies that allow to transfer into the industrial practice the Industry 4.0 approach for batch process monitoring. Within this dissertation, process monitoring is intended with a twofold meaning. On the one hand, process monitoring is required to early detect batches with an off-spec end-point product quality with the aim of minimizing the amount of produced off-spec batches. On the other hand, process monitoring is carried out for detecting anomalies in the process operating conditions, with the aim of troubleshooting the process, even if the end-point product is on specification.

First, a conventional data analytics technique is coupled to engineering knowledge for troubleshooting a semi-batch industrial process that is a bottleneck for the downstream sections. Data analytics is found decisive to identify an anomaly in the reactor safety interlock system that caused an increase in the time duration of batches in certain conditions. The interlock system is reconfigured and it is assessed that the intervention resulted in a 29% reduction in batch length, an 8% overall cycle time reduction and an 11% reduction in nitrogen consumption, entailing significant energy savings. The development of the data analytics model for this case study highlighted how batch alignment is an important and complex preprocessing step affecting the performance of the analysis. This finding was instrumental for developing techniques allowing the practitioner to easily carry out this step and completely avoiding it if not strictly necessary. In fact, artificial reduction of all batches to a common length (i.e., batch alignment) is usually carried out by trial and error, it is time consuming and it requires prior process knowledge.

A novel methodology for carrying out batch alignment in an automated manner is therefore developed. This methodology aims at preprocessing data for maximizing the performance of the model under development thanks to a surrogate optimization framework. The proposed method is completely process-agnostic, which enhances applicability to complex batch

processes. Also, in terms of computational times, it scales favorably with the calibration dataset size, and it is robust to normal variability in the data and to the size of the calibration dataset. An industrial fed-batch process for the manufacturing of a specialty chemical, and a simulated fed-batch process for the manufacturing of penicillin are used as test beds, and demonstrate that the proposed methodology has a superior performance than models built using other synchronization strategies.

When process monitoring is carried out with the aim of detecting anomalies in the process operating conditions one can use a methodology that does not need batch alignment, namely the assumption-free monitoring methodology proposed a few years ago in the literature. However, effective implementation of this methodology is challenging due to the lack of sufficient documentation and of a clear fault detection and diagnosis procedure. Hence, a set of detailed guidelines enabling the direct implementation of the methodology is developed together with a novel procedure for fault detection and diagnosis within this monitoring approach. Five datasets comprising batch processes from different industrial sectors are used for testing the methodology implemented according to the proposed guidelines, and proved that the proposed approach outperforms traditional process monitoring methodologies.

In conclusion, with this Dissertation an example of how process monitoring techniques are useful for improving the performance of industrial processes is provided, returning tangible results from an industrial point of view. The technique developed for automating phase partition and batch alignment performed better than other traditional and advanced techniques showing great potential for aiding the practitioner at developing models for end-point quality estimation without the need for prior process knowledge. Finally, the developed guidelines for the implementation of the assumption-free methodology proved to result in highly effective process monitoring schemes for operating conditions anomaly detection and diagnosis that can be implemented by the practitioner in a straightforward manner, overcoming the existing ambiguity and lack of information.

Table of contents

List of symbols

Acronyms

ARL	=	Average run length
BWU	=	Batch-wise unfolding
CAGR	=	Compounded annual growth rate
Cefic	=	European chemical industry council
COW	=	Correlation optimized warping
cw	=	Cooling water
DCS	=	Distributed control system
DTW	=	Dynamic time warping
EXT	=	Extension with mean values
FPR	=	False positive rate
GC-MS	=	Gas chromatography-mass spectroscopy
GHOPLS-CP	=	Generalized higher order PLS - canonical decomposition
HALS	=	Hindered amine light stabilizers
IV	=	Indicator variable
IVopt	=	Indicator variable optimization
IVopt	=	IV optimization
LDPE	=	Low density polyethylene
lps	=	Low pressure steam
LV	=	Latent variable
MPCA	=	Multiway PCA
MPLS	=	Multiway PLS
MS	=	Multisynchro
MSPM	=	Multivariate statistical process monitoring
NIPALS	=	Nonlinear iterative partial least squares
NOC	=	Normal operating conditions
o.o.l.	=	Out of limits
OECD	=	Organization for economic co-operation and development
OUT	=	Valve opening
P&ID	=	Process and instrumentation diagram
PARAFAC	=	Parallel factorization
PC	=	Principal component
PCA	=	Principal component analysis
Pensim	=	Penicillin simulator
PFD	=	Process flow diagram
PLS	=	Projection onto latent structures
PLS-DA	=	Projection onto latent structures and discriminant analysis

PV	=	Process value
PVC	=	Polyvinyl chloride
RBF	=	Radial basis function
RMSECV	=	root mean square error of cross validation
RMSEP	=	root mean square error in prediction
SOCMA	=	Society of chemical manufacturing & affiliates
SP	=	Set point
TPR	=	True positive rate
TR	=	Truncation
UV	=	Ultra violet
VWU	=	Variable-wise unfolding

Chapter 1

Motivation and literature review

The objective of this chapter is to provide the motivation of this Dissertation and the necessary background. The Industry 4.0 initiative is described and its impact on the chemical process industry is pointed out: furthermore, forecasts about how Industry 4.0 will impact the industry in the future are provided. Facts and figures about the specialty chemicals industry are then provided, the main features of the manufacturing of specialty chemicals are described and challenges and issues are pointed out. The industry critical task of process monitoring and how statistical process monitoring is enabling effective process monitoring based on process historical data is discussed, with a particular focus on batch process monitoring. Finally, the objectives of the research carried out in this project are provided together with a roadmap to guide the reader through the chapters composing this Dissertation.

1.1 Industry 4.0 in the process industry

"Industrie 4.0" (translated in English as "Industry 4.0") is a term originated from a German government initiative started in 2011 with the goal of improving the global competitiveness of the German manufacturing industry (Smit *et al.*, 2016).The term spread globally and it is now a worldwide adopted concept concerning "smart" and connected production systems designed to sense, predict and interact with the physical world in order to aid decision-making for real-time production support (Sirimanne, 2022).

The concept of Industry 4.0 refers to the fourth industrial revolution, after:

- the first industrial revolution begun at the end of the 18th century with the introduction of mechanical production plants based on water and steam power;
- the second industrial revolution started at the beginning of the 20th century with the symbol of mass labor production based on electrical energy;
- the third industrial revolution began in the 1970s with the first wave of automation based on electronic and internet technology.

The need for a fourth industrial revolution was triggered by several changes in the market and in the environment in which modern industrial companies operate. In particular the companies are required to (Lasi *et al.*, 2014):

- exhibit a high innovation capability by shortening the products time-to-market;

- being able to customize their products to their buyers' needs as the market has become more buyers driven;
- increase their flexibility in product development and production;
- become more efficient as raw materials and energy shortages and the related increases of prices have become more frequent. Furthermore, a higher social sensibility to ecological aspects requires a more intensive focus on sustainability.

The fourth industrial revolution is on-going thanks to the introduction of the smart factory concept, in which all objects are equipped with integrated processing and communication capabilities organized in so-called cyber physical systems (Lu, 2017), integrating both equipment and people in such a way that the individual skills and talents of everyone can be fully realized (Zheng *et al.*, 2018).

The core advantage of adopting the Industry 4.0 smart factory paradigm is that it is able to collect, distribute and integrate information of diverse nature dispersed across the supply chain and use it for enhancing safety, productivity and efficiency together with increasing the organizations sustainability and increasing the quality of products at the same time (Furstenau *et al.*, 2020; Reis and Kenett, 2018). Furthermore, studies suggest that the organizations environmental impact reduction thanks to Industry 4.0 can be quite significant (Oláh *et al.*, 2020). For these reasons, Industry 4.0 is considered to be a framework able to push the technological advancements required especially by certain industrial sectors for being able to comply to the increasingly stringent regulatory environment in terms of safety, efficiency and sustainability (Shang and You, 2019).

The boost in industrial performance through the application Industry 4.0 framework requires the development and the deployment of a synergistic combination of three factors (also called "key enablers"): data, technology and analytics (Reis and Gins, 2017).

The computational and data storage capabilities of a modern process plant allow it to log more than 90,000 tags at any time (Chiang *et al.*, 2022). The change occurred in the amount of registered data influenced not only the number of the variables, but also the nature of the registered data. In fact, a wide variety of sensor types is able to provide chemical signals (e.g. spectra), physical signals (e.g. pressures, temperatures, flows), images, sounds and so on (Ferrer, 2020).

It is forecasted that Industry 4.0 will continue to influence the industrial development over the next 10 years thanks to a number of emerging technologies. In particular, an even increased speed in the network connections thanks to new networking technologies should allow even larger amounts of data to be transferred from the equipment to data storage facilities, where these data are used to enable through machine learning approaches self-learning capability, thus extreme flexibility, of the smart factory, guaranteeing robust production, high occupational safety, energy efficiency and a high degree of resource conservation, together with the ability to respond quickly to market volatility through a revolutionized approach to production

planning. The effectiveness of this approach is further increased by integrating industry requirements for data transparency, reliability and trustworthiness (Kagermann and Wahlster, 2022). The development of analytics techniques able to cope with the characteristics of data originated in an Industry 4.0 framework will be paramount to enable better decision making, thus, boost industrial performance.

1.2 Specialty chemicals and polymer additives

In the chemical process industry, raw materials are converted into products using energy, for other industries and consumers. Around the 85% of the total production is taken by a limited number (about 20) of simple chemicals called base chemicals. The conversion of base chemicals produce around 300 intermediates. Both base chemicals and intermediates are classified as bulk chemicals (Moulijn *et al.*, 2014). The specialty chemicals (also called "performance chemicals" or "specialties") segment of the chemical process industry consists of businesses that convert intermediates into active ingredients for use in end products (Bonvin *et al.*, 2006). With each of the abovementioned steps, the complexity of the molecules becomes larger and the added value of the chemicals becomes larger.

Specialties can be either single molecules or formulations that influence the performance of the end product (American Chemistry Council, 2022). Unlike commodity chemicals, that may have a wide array of different applications, a specialty only has one or two core applications (SOCMA, 2023). Examples of specialty chemicals include polymer additives, adhesives and sealants, catalysts, coatings, electronic chemicals, cleaners, water management chemicals (American Chemistry Council, 2022). Specialty chemicals are instrumental in our everyday life and are instrumental in producing many of the foods we eat, medicines we take and the clothes we wear (Guisinger and Ghorashi, 2004).

Despite being hit by unprecedented challenges, such as COVID-19 and the Russo-Ukrainian war (McKinsey, 2020; OECD, 2022), the economic outlook of the global specialty chemicals market is positive, with a moderate forecasted growth from 738.23B$ in 2022 to 998.94B$ in 2028, with an average compound annual growth rate (CAGR) of 5% (BusinessWire, 2023), although the global chemical market is expected to be weakening in all countries (BASF, 2022). Although being produced in low volumes by definition, specialty chemicals represented 28% of the European chemicals market share in 2021, exhibiting a moderate increase of its share from the 26% recorded in 2010 (Cefic, 2023).

Polymer additives are a relevant subset of specialty chemicals used to aid or facilitate polymer processing or to enhance, extend or modify the final properties of polymer products, some of these additives include UV absorbers, antioxidants, blowing agents, plasticizers, lubricants, flame retardants. Polymer additives are used in many industrial sectors, such as automotive, building and construction, electronics, packaging.

Although representing a small portion of a customer's total cost, specialty chemicals, such as polymer additives, are essential to the productivity and the performance of a product (Allen and Edge, 2021a, 2021b).

1.3 Features of the production of specialty chemicals

Specialty chemicals are high value-added, low volume products. One of their most important features is the existence of a great variety of specialties, with new products continuously emerging. Often, significant fluctuations in the demand exist.

These characteristics of specialties make the construction of plants dedicated to single products costly. Hence, usually specialty chemicals are produced in multiproduct or multipurpose plants where multiple products are produced using the same pieces of equipment, usually in batch or semi-batch operating mode (Moulijn *et al.*, 2014).

Batch processes are characterized by the cyclic repetition of a recipe, i.e., a sequence of elementary finite duration processing steps (such as: charge, heat up, stir, react, cool down, hold, discharge). A set of operating conditions characterizes each of the processing steps and is typically triggered by the occurrence of events (e.g., enough reactant has been loaded, a certain temperature value is reached, agitator torque exceeds a threshold).

Batch processes play a dominating role over other production modes (e.g. continuous) in the production of specialty chemicals due to the following desirable characteristics (Rato *et al.*, 2016):

- operational flexibility;
- production scalability;
- ease of setup and operations.

Specialty chemicals are produced in plants composed of a complex interconnected network of batch process units: many of these units can be used both for their main scope, but also as buffer tanks in case the downstream units are busy with the manufacturing of previously produced batches. Batch processes can be operated easily and with flexibility, and due to these characteristics, they are usually preferred when processes with complex chemistry are involved, such as it is usually the case with specialty chemicals, and with limited mechanistic knowledge of the process.

This set of advantages in adopting the batch manufacturing mode comes at a cost. Adjusting the length of processing steps is an approach to accommodate variability in the raw materials, operating conditions, status of the equipment and of the utilities. However, if the length of such steps is increased, it may reduce the overall productivity of the process. Furthermore, maintaining a batch process in the right operating conditions for a longer amount of time may require a larger use of utilities, reducing the efficiency of the process itself. As the time required

for producing each batch may vary, planning and scheduling is more complex and may lead to challenges in meeting delivery deadlines and managing inventory levels effectively.
Product quality is usually measured only at the end of the batch on the end-product, making it harder to introduce appropriate corrections to the recipe in a timely manner, in order to obtain on-spec products.
Proper monitoring and control schemes need to be deployed in order to mitigate the risks induced by the larger variability allowed to enter the process in batch processing (Rendall *et al.*, 2019).

1.4 Statistical process monitoring in the process industry

In recent years, the process control discipline made huge progresses with the advent of computer control of complex processes. Regulatory control (i.e. opening and closing valves), a task that used to be carried out by human operators, is now routinely performed automatically with the aid of computers.
However, higher level control tasks, remain largely a manual activity performed by human operators (Venkatasubramanian *et al.*, 2003a). The most important of such tasks is detecting, diagnosing and responding to abnormal events (also called faults) in a process, or, stated in a more concise manner, process monitoring (MacGregor and Cinar, 2012). The number of process variables observed every few seconds in a modern production plant may lead human process supervisors to information overload if observed separately. Furthermore, the process measurement may be insufficient, incomplete and unreliable due to a variety of causes such as sensor failures. In such a complex environment human operators tend to make erroneous decisions and take wrong actions. Industrial statistics show that human errors are the main cause of industrial accidents. Process monitoring technology has been developed for assisting operators in ensuring product quality and safe operations (Ji and Sun, 2022).
Process monitoring consists of four activities (Chiang *et al.*, 2001):

- fault detection;
- fault identification;
- fault diagnosis;
- process recovery.

There is no standard terminology for these procedures because they vary across disciplines, hence in this Dissertation we will use the terminology described in the remainder of the current Section (Raich and Çinar, 1996).
Fault detection consists in determining whether a fault has happened. In particular, detecting faults in a timely manner may provide useful information enabling to take action to avoid serious process upsets or accidents. Fault identification consists in identifying the observed variable most relevant to diagnosing the fault. The purpose of this step is to identify the plant

section that is most pertinent to the observed fault, so that the effect (or the cause) of the fault can be eliminated. Fault diagnosis consists in assessing the type, location, magnitude and time of the fault. At last, process recovery (or intervention) consists in removing the effect of the fault. In a process monitoring scheme either all or a subset of these 4 procedures can be implemented. Furthermore, it is not needed to automate all four procedures, as they are intended to support plant engineers and operator to recover normal operating conditions (NOC).

In order to be able to detect deviations from the NOC, a model representing such conditions must be developed. Although the typical process modelling approach in the chemical and process engineering context is based upon first-principles knowledge on the process, this approach is usually expensive and time consuming, and may require extensive experimental effort. On the other hand, thanks to process digitalization, the wide availability of process data rendered possible to use empirical (or data-driven) models for the study of process systems (Venkatasubramanian *et al.*, 2003b). This approach is particularly useful in cases where first-principles knowledge on the system is scarce, which often happens in the production of specialty chemicals. Industrial data are often characterized by a high dimensionality and are often affected by spatial and serial correlation, multicollinearity, noise and the presence of missing data. Multivariate statistical models, in particular latent variable modelling methodologies such as the principal component analysis (PCA; Jackson, 1991), the projection onto latent structures (PLS; Geladi and Kowalski, 1986) and its extension for classification problems, projection to latent structures – discriminant analysis (PLS-DA, Barker and Rayens, 2003) are convenient frameworks for process modelling in this context, as they summarize the physico-chemical phenomena underlying a process with a reduced number of latent variables (LVs), handling multicollinearity, noise and missing data and extracting hidden information in the correlation pattern between process variables altogether (Eriksson *et al.*, 2006).

Both PCA and PLS(-DA) are concerned with representing the correlation structure of the data through linear combinations of the original variables. However, PCA is used when a single block of data is analyzed (e.g. process data), while PLS(-DA) is the methodology adopted when the relationship between process and quality data needs to be investigated (Wold *et al.*, 2001).

Once the statistical characteristics of the NOC batches are captured by a latent variable model, limit sensing and discrepancy detection are deployed, calculating statistical limits on the multivariate Hotelling's T^2 statistic and the Q-statistic for fault detection. These two statistics give an overall account of abnormal situations in a complementary manner (Qin, 2012), the former describes the distance of an observation from the NOC conditions, while the latter describes how accurately the observation is described by the model itself, further information on this topic will be given in Chapter 3.

1.4.1 Multivariate statistical process monitoring for batch processes

Developing proper techniques for batch process monitoring is important both to achieve a consistent product quality from the batch process itself, and for running the process in safe and controlled conditions (Chiang *et al.*, 2006). Modelling batch processes however is a very challenging task due to various factors: their duration is finite and variable, their nonlinear and irreversible behaviour, lack of mechanistic and fundamental models, sensor inaccuracy, existence of constraints; unmeasured disturbances.

The data collected online from batch processes are multivariate in nature, but are also affected by nonlinearity, auto and cross-correlation. Furthermore, they are time varying, for this reason along the two dimensions of data usually considered in continuous processes (observations and variables), a third dimension must be considered for batch processes: time.

Thus, the data collected from a batch process can be collected in a 3 dimensional tensor with batch runs on the first dimension, process variables on the second dimension and time on the third dimension (Kourti, 2003).

A variety of methodologies have been proposed for batch process monitoring, however, the selection of the best methodology for a given application is often an ill-posed problem that practitioners must face. These methodologies can be classified into three main classes, in order of complexity, each with advantages and disadvantages (Rendall *et al.*, 2019):

- feature-oriented methods;
- linear time-resolved methods;
- nonlinear time-resolved methods.

In the following subsection each category will be described together with its advantages and disadvantages.

1.4.1.1 Feature-oriented methods

Feature-oriented methods are based upon transforming batch profiles into a set of features capturing monitoring-relevant characteristics of the process variables trajectories. The underlying principle of such methods is borrowed from classical pattern recognition analysis, and it consists in finding a transformation from the measurement space to the feature space, where the identified feature space contains all the necessary information to successfully conduct the process monitoring task. In literature, different classes of features have been proposed: simple descriptive features for the batch variables (e.g. maximum, minimum, phase duration), statistical features (e.g. mean, covariance, skewness, kurtosis), features based on wavelet decomposition (Rendall *et al.*, 2017a).

The main disadvantage of feature-oriented methods is that they have been only applied in an offline manner, since feature can be computed only when the whole variable time trajectories are known (Rendall *et al.*, 2017b).

1.4.1.2 Linear time-resolved methods

Linear time resolved methods are a step up in the complexity scale with respect to feature-oriented methods. Their advantage is that they maintain time resolution, i.e., they can inherently describe not only if a fault occurred, but also when the fault occurred, at the cost of a relevant increase of the model parameters with respect to feature-based models.

One approach to linear time resolved methods is to model the 3 dimensional tensor of batch process data directly, such as in PARAFAC, PARAFAC2 and Tucker3 methods (Louwerse and Smilde, 2000), the main disadvantage of the application of these methodologies is that the algorithms calculating the model are often inefficient and their computational burden is high. Furthermore, they tend to be sensitive to noise (Amigo *et al.*, 2008; Tian *et al.*, 2018).

Other approaches based on the multiway versions of PCA and PLS consider the 2-way unfolding of the batch data tensor.

Unfolding the batch data tensor along the batches direction gives place to the so-called batch-wise unfolding (BWU; Nomikos and MacGregor, 1995). This method captures the dynamics of the process, summarizing the information of the data with respect both to variables and their time variation, while the major nonlinearities in the process data can be removed by subtracting the average trajectory from each variable (Nomikos and MacGregor, 1994). Furthermore, BWU is the most logical method for modelling differences among batches, especially in the common situation where a single quality measurement at the end of the batch is available for each batch (Golshan *et al.*, 2010).

Although being an effective method for a batch-to-batch monitoring strategy, it has some relevant drawbacks that need to be managed. In order to use BWU MPCA and BWU MPLS(-DA) for batch monitoring, the entire history of a batch should be available while monitoring a batch in real time in order to be able to complete the 2-dimensional batch data matrix.

This aspect can be solved by imputing the future history of a batch (Nomikos and MacGregor, 1995a), assuming that the BWU methodology works appropriately if at least 10% of the batch evolution is known. Furthermore, a crucial aspect when using BWU MPCA and BWU MPLS(-DA) is that, in order to be applied, batch data need to be aligned, i.e., the time trajectories of all batches need to have the same number of time points. In order to achieve this, a wide number of methodologies have been proposed in literature, from simply truncating batches to the length of the shortest or extending them to the length of the longest (Lakshminarayanan *et al.*, 1996; Rothwell *et al.*, 1998), to more advanced methodologies, such as the indicator variable method (IV; Nomikos and MacGregor, 1995), to methodologies borrowed from other research fields, such as dynamic time warping (Kassidas *et al.*, 1998) and correlation optimized warping (Tomasi *et al.*, 2004) and their offspring (José M. González-Martínez *et al.*, 2014). In order to render possible to align a batch with some of these techniques, it is needed to partition a batch in phases, i.e. time windows wherein the measured variables have similar correlation structure

(García-Muñoz *et al.*, 2003), however this task is challenging as phases do not generally correspond to events along a batch.
A different approach consists in unfolding the batch data tensor along the variable direction, obtaining the so called variable-wise unfolding (VWU; Wold *et al.*, 1998).
It is in principle possible to apply the VWU methodology to batch data without batch alignment and missing data imputation, however it is crucial to apply some form of alignment in order to not only have batch trajectories that are equal in length, but also to synchronize the key events defining the normal process pace (González-Martínez *et al.*, 2014). Even though VWU is a simpler methodology than BWU for batch processing MSPM, it does not consider the time order of the batch samples and it forces the correlation structure to remain constant along the entire batch. For these reasons, the standard version of VWU is deemed to have inferior performance in batch process monitoring with respect to BWU.
A methodology overcoming some of the assumptions of both BWU and VWU based MPCA and MPLS monitoring techniques is the assumption-free monitoring technique (Westad *et al.*, 2015). This technique is a methodology that can accommodate uneven batch lengths, unknown initial batch (absolute) time, phase changes and uneven residence time by inherently estimating the so-called "relative time", a measurement of progress of the underlying chemical, biological and physical phenomena along the process. The estimation is carried out through gridding of a portion of the latent variable space and building a common batch trajectory representing the NOC.

1.4.1.3 Nonlinear time-resolved methods

The higher level of complexity in the methodologies for batch process monitoring is represented by nonlinear time-resolved methods. These methodologies are able to describe effectively nonlinear relations in the datasets. They are the most flexible among the described methods and are often based on kernels mapping samples into high dimensional non-linear spaces where the original process variables are implicitly modelled (Rendall *et al.*, 2019). Even though these methodologies are more powerful than linear time-resolved methodologies, fault diagnosis is rendered difficult by the nonlinear feature extraction performed through the kernel transformation (Pilario *et al.*, 2020).

1.5 Objectives of the research

Despite a high academic and industrial interest in batch process performance monitoring, not only for real-time fault detection and diagnosis, but also for data-driven process optimisation through retrospective analysis of process data, its effective implementation in modern industrial batch processes is challenging.

The pre-processing of the batch process data is crucial for the successful implementation of a data-driven batch process monitoring scheme, but the two most relevant steps, phase partitioning and batch alignment, are performed separately and a priori with respect to modelling.

For fault detection and diagnosis, phase partitioning and synchronisation can be avoided by using techniques that do not require them, such as the assumption-free monitoring technique. However, a challenge for its implementation is the scarcity of available information, and the complete lack of certain procedures, such as fault detection and diagnosis. In quality related process monitoring, when performing batch endpoint product quality estimation, a methodology that avoids phase partitioning and batch alignment has not yet been developed, so quality related process monitoring still requires extensive pre-processing, exposing the practitioner to the risk of human error in selecting the most appropriate pre-processing techniques.

The objective of this project is the development of Industry 4.0 approaches based upon multivariate statistical techniques for batch process monitoring that reduce subjectivity in the preprocessing step in the development of batch process monitoring by systematizing and automating it, and by developing guidelines and complete existing methodologies for batch process monitoring that do not require phase partition and batch alignment.

The innovative contributions aiming at this objective that can be found in this dissertation are the following:

- **Assessment of the benefits of the application of batch process monitoring techniques** on a real and complex (multi-unit) industrial process, with a particular focus on the impact of applying Industry 4.0 monitoring techniques to the manufacturing process for the production of speciality chemicals. The focus of the research is to assess whether troubleshooting a real industrial process with data-driven modelling techniques can improve the performance of the process itself. The use of this modelling approach is made possible by the availability of large amounts of data, thanks to advances in sensors and data storage capabilities triggered by Industry 4.0. In particular, reducing the environmental impact of a process by reducing its consumption of utilities and energy can also significantly reduce the operating costs of the process. In addition, it will be investigated whether it is possible to achieve a reduction in the process cycle time, thus allowing an increase in the productivity of the process itself.
- **Development of automated approaches for batch alignment and phase partition for end-point quality estimation**. The development of soft sensors and classifiers for end-point quality assessment in batch processes is of great importance to reduce the burden on the quality control laboratory and at the same time reduce the time required to perform laboratory assays to assess end-point product quality. In batch processes, the development of such models typically requires the application of complex pre-

processing, such as batch alignment and phase partitioning, often in a subjective and sub-optimal manner, resulting in the development of sub-optimal models. Furthermore, the application of existing phase partitioning and batch alignment techniques usually requires some process knowledge for their application. The research effort will focus on a methodology that allows the practitioner to automatically obtain the batch alignment and phase partitioning that maximises the performance of the final model. The final objective is to develop a robust phase partitioning methodology based on the correlation between process variables, to develop an automated approach for batch alignment based on the IV technique, which has proven to be simple and effective, and to optimise these steps through the use of a surrogate to improve the computational efficiency of the methodology itself. The methodology will be entirely data-driven and must not require prior information on the process under investigation to be applied. The performance of the developed method with respect to existing advanced batch alignment methods will be evaluated both in terms of model loss function and computational burden.

- **Development of guidelines for phase partition and batch alignment-free methodologies** for further reducing the time expense for the development of data-driven models for batch process monitoring. The batch alignment step is a time-consuming pre-processing step that introduces distortions and artifacts to the process variables time trajectories, hence, eliminating the need for carrying it out allows for a quicker and improved model development. Currently, a multivariate statistical methodology not requiring batch alignment and phase partition has been proposed in the literature. Although it attracted industrial interest, as it is implemented in commercial software, it did not attract much academic interest. This technique can be applied on time-resolved variables, such as time profiles of process variables. The main objective is filling in the existing gaps in the information required for model implementation and to develop extensive guidelines for the practitioner in order to allow an easy implementation of the methodology. The guidelines will be completed by a discussion on the techniques that must be adopted for carrying out fault detection and diagnosis through a phase partition and batch alignment-free methodology and with the development of novel tools for carrying out these activities. The methodology implemented with the proposed guidelines will be compared with a traditional process monitoring methodology on different case studies for assessing its performance in terms of detection strength and speed and in terms of fault diagnosis effectiveness.

1.5.1 Dissertation roadmap

The Dissertation is organized following the research objectives illustrated in the previous section. A schematic roadmap is shown in Figure 1.1.

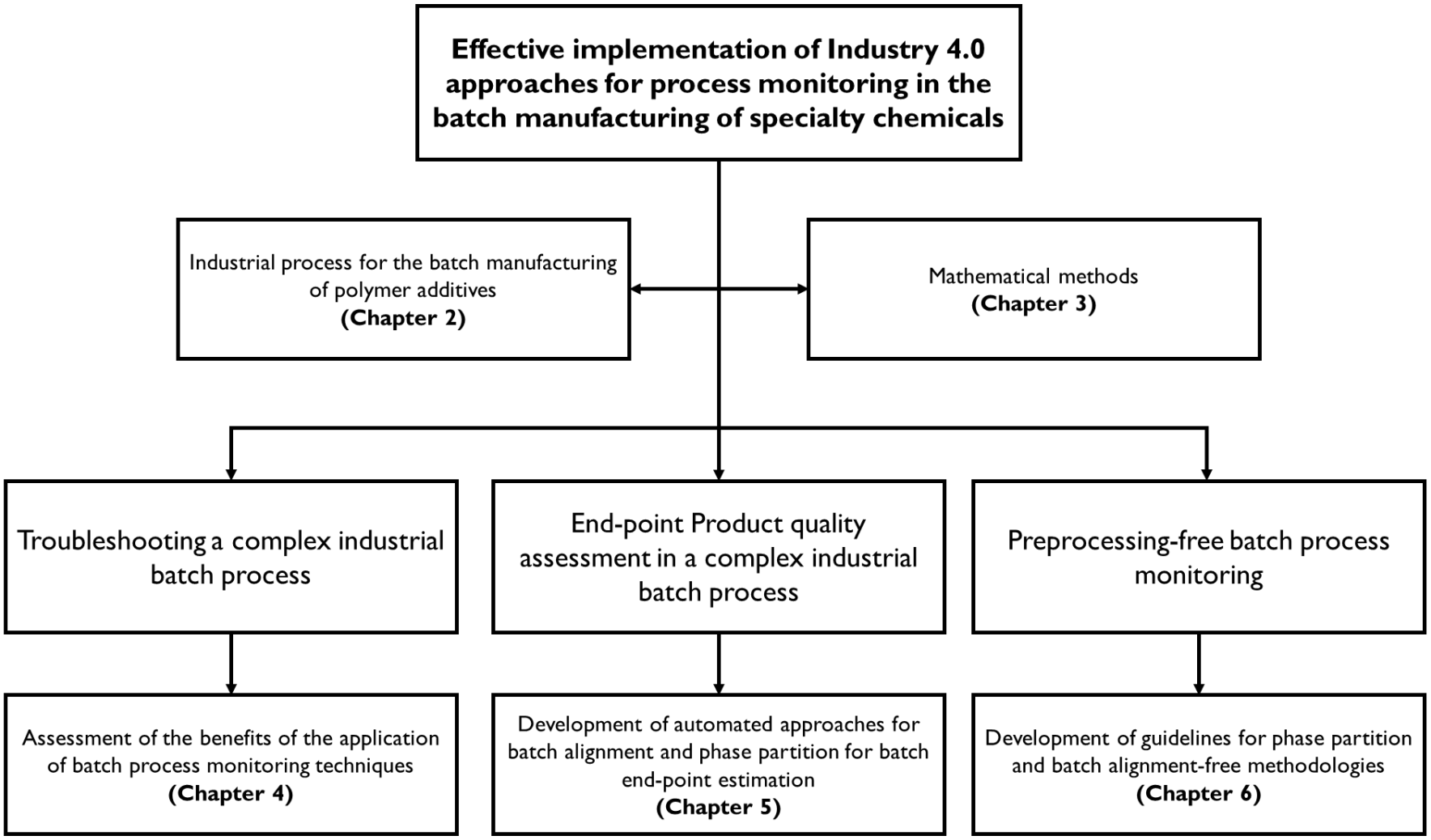


Figure 1.1: *Dissertation roadmap*

Chapter 2 contains a contextualization of the industrial process in the specialty chemicals sector of the chemical process industry where the challenges present in the manufacturing of specialty chemicals will be tackled and the process under study will be described.
In Chapter 3 a brief review of the mathematical tools exploited throughout the Dissertation is provided with a particular focus on multivariate statistical methodologies, together with a description of the most relevant phase partition and batch alignment methodologies used in data preprocessing for batch process monitoring.
In Chapter 4 the benefits of the application of a process monitoring scheme to a real industrial process manufacturing specialty chemicals will be shown. A classical approach to process monitoring will be adopted and it will be shown that it is able to improve safety and energy efficiency in a process, as well as detecting an unknown and undetected fault worsening the productivity of the process. The limitations of using a classical approach will be pointed out.
In Chapter 5 an automated approach to phase partition and batch alignment is presented for a process-agnostic, human error free preprocessing step in batch end-point quality estimation for process performance monitoring.
Chapter 6 provides guidelines for the application of a process monitoring methodology not requiring phase partition and batch alignment. The methodology is further completed with novel approaches for fault detection and diagnosis in assumption-free monitoring. The Dissertation is concluded with final remarks.

Chapter 2

Industrial process for the batch manufacturing of polymer additives

In this Chapter the properties of the final product of the process under investigation, hindered amine light stabilizers (HALS), are discussed. A description of the process is provided together with a simplified process flow diagram of the overall process and several piping and instrumentation diagrams representing the most relevant process units. A description of the recipe of the most relevant units is provided together with information on the chemistry of the process. After that, a description of the data acquisition process is given and completed by a discussion on the properties of the extracted process and product quality data. Eventually, the advantages and disadvantages of the application of an Industry 4.0 approach to the production of specialty chemicals is delivered.

2.1 Hindered amine light stabilizers

Hindered amine light stabilizers (HALS) are a class of polymer additives used to protect materials, particularly polymers, from degradation caused by exposure to light (Rabek, 1990). Their typical molecular structure is depicted in Figure 2.1.

HALS, depending on their chemical structure, can exist as liquids or solid powders. HALS work by inhibiting the degradation process caused by ultraviolet (UV) radiation. When exposed to UV light, polymers can undergo photochemical reactions that lead to chain scission, crosslinking or free radical formation, ultimately weakening the material and causing it to lose its physical and mechanical properties (Guillet, 1972).

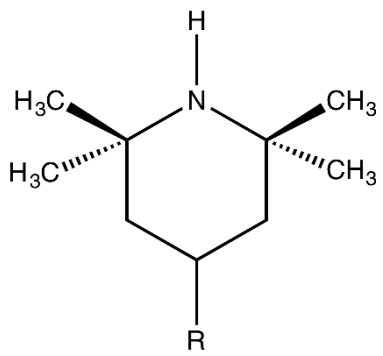


Figure 2.1 *Structure of a typical HALS (adapted from Rabek, 1990)*

HALS act as free radical scavengers, intercepting the harmful free radicals formed during photodegradation. By neutralising these radicals, they prevent further degradation reactions, extending the life of the material and increasing its stability. One of the key advantages of HALS is their ability to act as effective and long-lasting stabilisers. Unlike other light stabilisers, they do not undergo significant chemical changes during the stabilisation process, allowing them to provide extended protection over time (Carlsson *et al.*, 1984).
HALS can be incorporated into materials during processing or applied as a surface coating (Schaller *et al.*, 2009). They are effective in a variety of polymer materials including polyolefins, polyurethanes, PVC and others (Cui *et al.*, 2020; Mousavi-Fakhrabadi *et al.*, 2022).

2.2 Manufacturing process

Specialty chemicals are high value-added products often produced through batch processing. Batch processes are characterized by the cyclic repetition of processing steps (e.g. reactant charge, heating/cooling, reaction, mixing, product discharge) through a so-called recipe.
The most relevant properties of batch processes are represented by their *i)* operational flexibility; *ii)* production scalability and *iii)* ease of setup.
Operational flexibility is enabled by a combination of factors. Batch processes can be operated with minimal supervision when a sufficient degree of automation is implemented in the production plant. Furthermore, the recipe can be adjusted to accommodate changes in raw materials, state of equipment and utilities. Production scalability is possible since batches are treated independently one from the other, hence it is possible to change the number of batches produced over time without changing the process, within certain boundaries. Furthermore, scale-up is straightforward as batch reactors of any required scale are easily procurable. Due to these desirable characteristics, batch process units are usually included in a complex, interconnected network of process units as flexibility in the connections between process units is required when different products must be produced with the same units.
A batch process unit can be usually employed both for his main scope (e.g., reaction, separation) but also as a buffer tank if downstream units are still busy with the manufacturing of a previous batch or for productivity reasons. For quality control reasons, information on product quality is required. Usually, a sample for a reduced number of batches is taken at the end of the batch from process units for laboratory analysis. However, this is true only for a small number of process units that are considered relevant for product quality. As performing laboratory assays is usually expensive and time consuming, process units considered less important for product quality are operated without control on product quality. Hence, for some intermediate products laboratory protocols for quality control have not been developed at all.
In Figure 2.2 a representation of the process under investigation is shown. Reactants and products are designated with a letter for protecting the confidentiality of the industrial data.

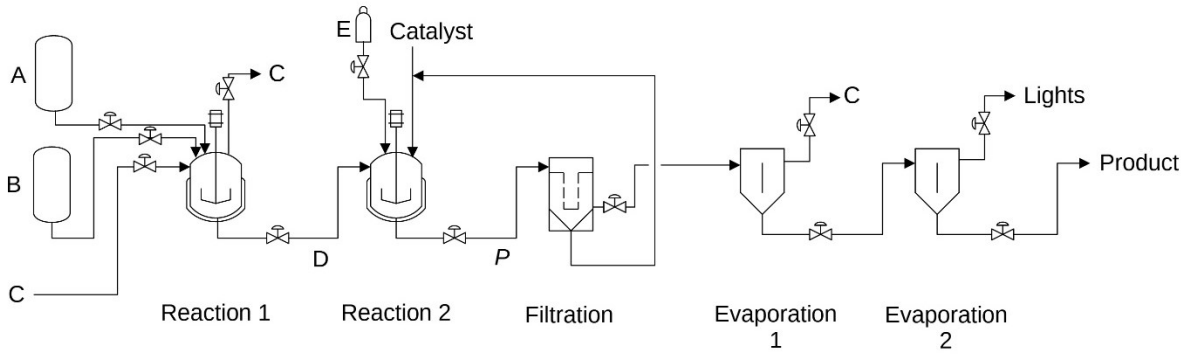


Figure 2.2 *Simplified process flow diagram of the process under investigation*

The manufacturing process starts with reactant A and B loading into Reactor 1, where a semi-batch process is carried out. Vacuum is applied to the system, and the following liquid-phase, thermally activated, exothermic reaction takes place:

$$2\mathrm{A} + \mathrm{B} \rightarrow 2\mathrm{C} + \mathrm{D} \quad , \tag{2.1}$$

where A and B are two different amines, C is an inorganic byproduct. and D is an imine.
The process carried out in Reactor 1 is highly automated and consists of the following steps:

1. setup: the system is set up for a new run;
2. reactant loading: liquid reactants A and B are loaded into Reactor 1 from their respective storage tanks in stoichiometric amounts. A small amount of C is also loaded into the reactor; the reason for this is to speed up the initial reaction phase. The reactor is heated up with low-pressure steam through a jacket, to reach the temperature required to carry out the reaction;
3. reaction: vacuum is applied to the system in two steps: a faster pressure decrease is applied first, down to an assigned pressure; then, pressure is further slowly decreased to the pressure value required by the reaction. Once the reaction conditions are met, byproduct C is released as a vapor, and it is removed from the reactor;
4. nitrogen blanketing: when the required amount of evaporated water is obtained, vacuum is broken, and the reactor is blanketed with nitrogen;
5. product discharge: reaction product D (liquid) is discharged from Reactor 1 and fed to the subsequent processing step.

The product of Reactor 1 is then fed to Reactor 2 together with a solid catalyst. In Reactor 2 a fed-batch process is carried out and the main product *P* of the process is obtained from the following catalytic reaction

$$\mathrm{D} + \mathrm{E} \xrightarrow{\text{Catalyst}} \mathrm{G} \quad , \tag{2.2}$$

where D is a liquid reactant, E is a gaseous reactant and G is the desired species (a HALS). Product *P* is mainly made of G, traces of unreacted B and C and other subproducts. Furthermore, it still contains the solid catalyst.

The manufacturing recipe for Reactor 2 is complex and it can be summarized by the following finite-length operating steps:

6. Reactor 2 is set up for a new batch.
7. Reactant B and catalyst are loaded into Reactor 2.
8. Reactor 2 is blanketed with nitrogen.
9. Reactant D is fed to Reactor 2 and pressurizes it until an assigned pressure is reached; after that, the feed is stopped, and the reaction is allowed to proceed for an assigned amount of time. The profile through which B is fed depends on several factors and is quite complex, resulting in a very strong variability of this phase.
10. Reactor 2 is vented.
11. Reactor 2 is blanketed with nitrogen.
12. Product *P* is discharged from Reactor 2 to downstream separations.

Product *P* is then fed to a filtration step where the solid catalyst is separated and recycled for a definite number of times to Reactor 2, after the maximum number of recycles is reached, the catalyst is replaced with fresh catalyst. While the liquid phase is fed first to a continuous section of the process, Evaporator 1 where the reactant C that was not removed in Reactor 1 is separated from the main product, and then to Evaporator 2 where light byproducts are separated from the product, which now respects the specifications required for being fed to subsequent processes for the synthesis of HALS light stabilizers.

2.3 Production plant

A more detailed process flow diagram for the process described in Section 2.2 is shown in Figure 2.3.

Reactant A is loaded from storage tank T1 (where it is stored after its in-house production) and Reactant B is loaded from storage tank T2 (outsourced) into reactor R1 (reactor volume: 7.5 m^3) where reaction 1 process takes place. The product of R1 is then discharged either to T3 (buffer tank volume: 6 m^3) or to T4 (buffer tank volume: 6.7 m^3). The batch then can be sent in one of 6 multiproduct reactors (R2 to R7): if the batch is stored in buffer tank T3 it is sent either to R2 or R3, if it is stored in buffer tank T4 it is sent to R4, R5, R6 or R7. These reactors carry out the reaction 2 process.

R2 and R3 feed one buffer tank, T5 (volume: 7.5 m^3), feeding filter F1, while reactors from R4 to R7 can feed 2 different buffer tanks, T6 (volume: 7.1 m^3) or T7 (volume: 14.4 m^3) feeding filter F2 and F3, respectively.

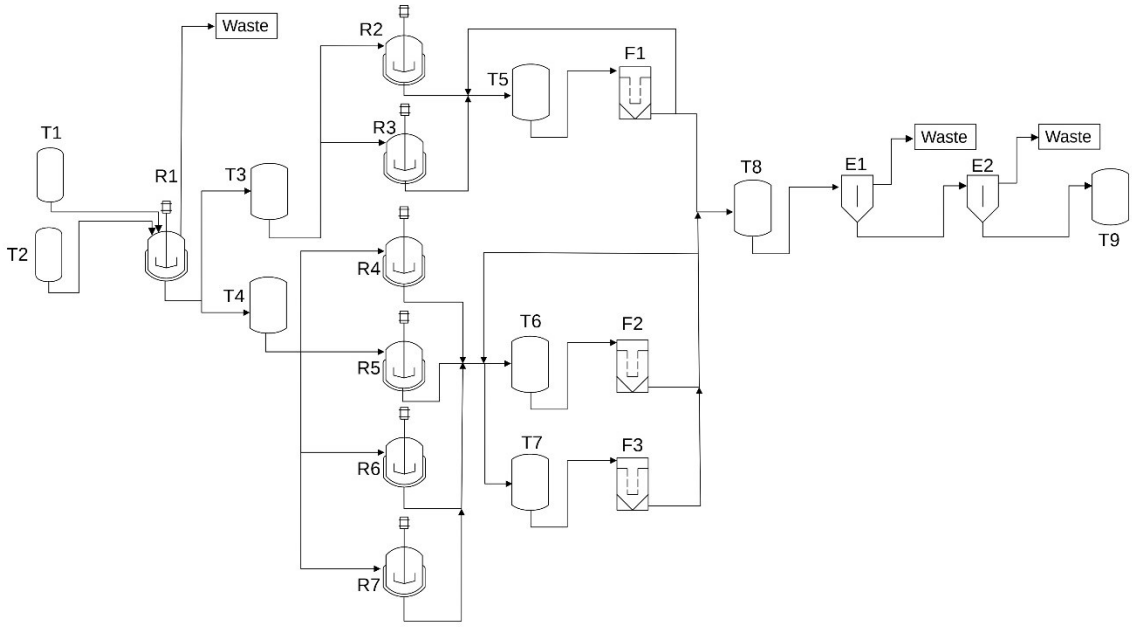


Figure 2.3 *Process flow diagram of the process under investigation.*

The filters feed a buffer tank, T8 (volume: 14.4 m^3) that is used as the feed for the continuous separation section constituted by evaporators E1 and E2. At the end of the plant, the final product tank T9 (volume: 100 m^3) stores the product of the process.

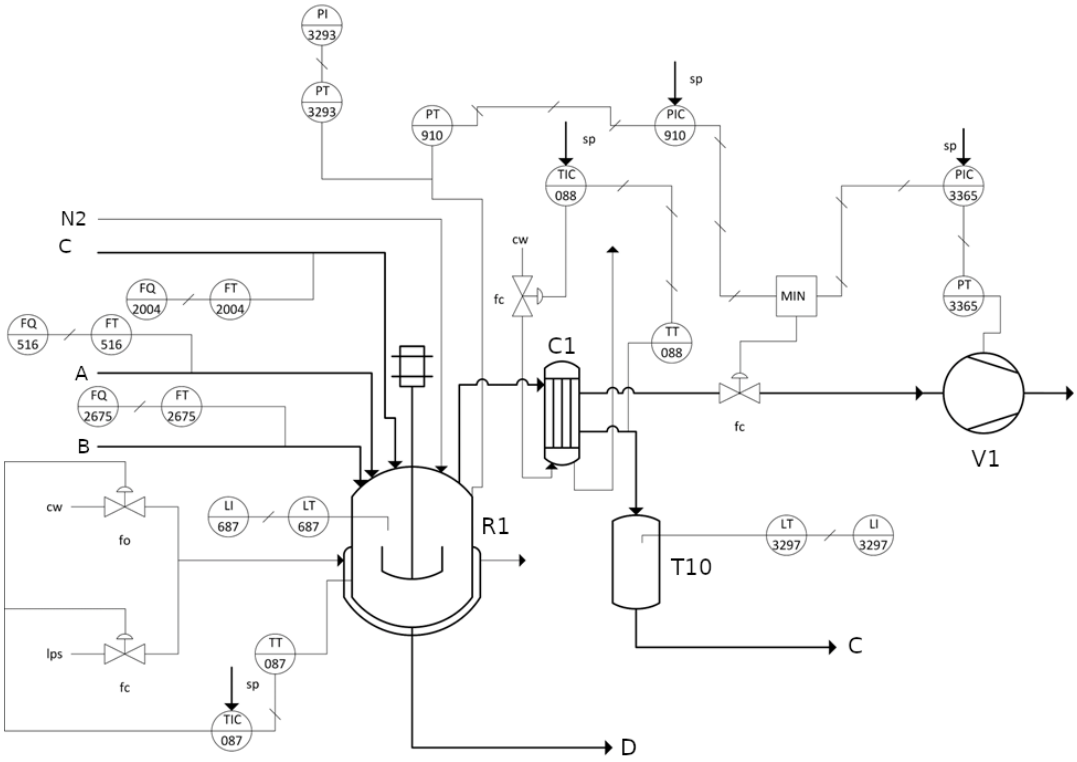


Figure 2.4 *Piping and instrumentation diagram of reactor R1*

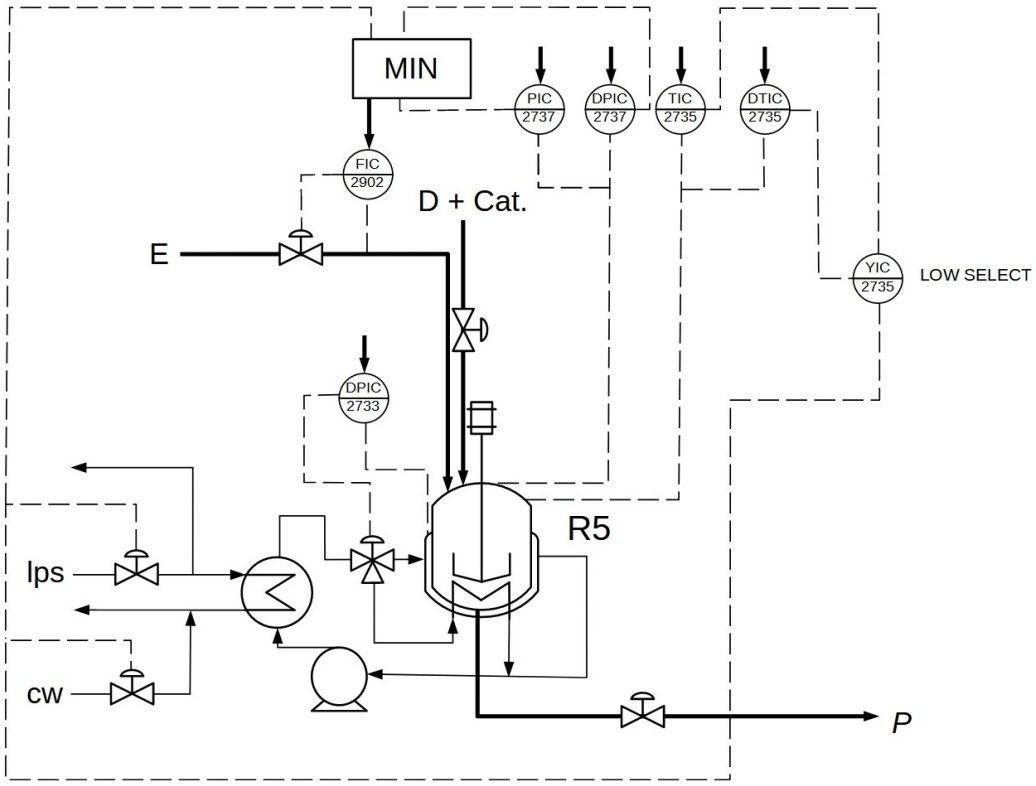


Figure 2.5 *Piping and instrumentation diagram of reactor R5*

A detailed piping and instrumentation diagram of reactor R1 and reactor R5 are shown in Figure 2.4 and Figure 2.5, respectively.

2.4 Data acquisition

The data produced in the plant are managed by a distributed control system (DCS) software produced by ABB (ABB 800xA). The DCS collects uncompressed data from the connected field sensors and its storage capacity is enough to store the last three months of data only.
Data older than 3 months are sent to a data management system (data historian) PI (produced by OSIsoft, LLC), where they are collected and elaborated. The data stored in PI are compressed.
It is possible to access PI data in a number of ways:

- via an extension for Microsoft Excel developed by OSIsoft;
- via an in-house developed MATLAB script;
- via a commercial graphical user interface (GUI), PARCview (produced by dataPARC);
- via the commercial TrendMiner web based platform;
- via the OSIsoft proprietary software PI Process Book and PI Vision.

All these different software solutions allow to extract data as a Microsoft Excel file for further elaboration and analysis.

2.4.1 Process variables

For each of the process units depicted in Figure 2.3, a number of online measurements varying from 20 to 70 is available in the data historian. The very large number of variables available makes listing all the variables not useful for the reader, hence the used variables will be listed in the following Chapters. The available measurements typically include process values (PV), setpoints (SP) and valve openings (OUT). In some cases, the same variable is measured by two or more sensors, as redundancy is desirable for safety reasons.

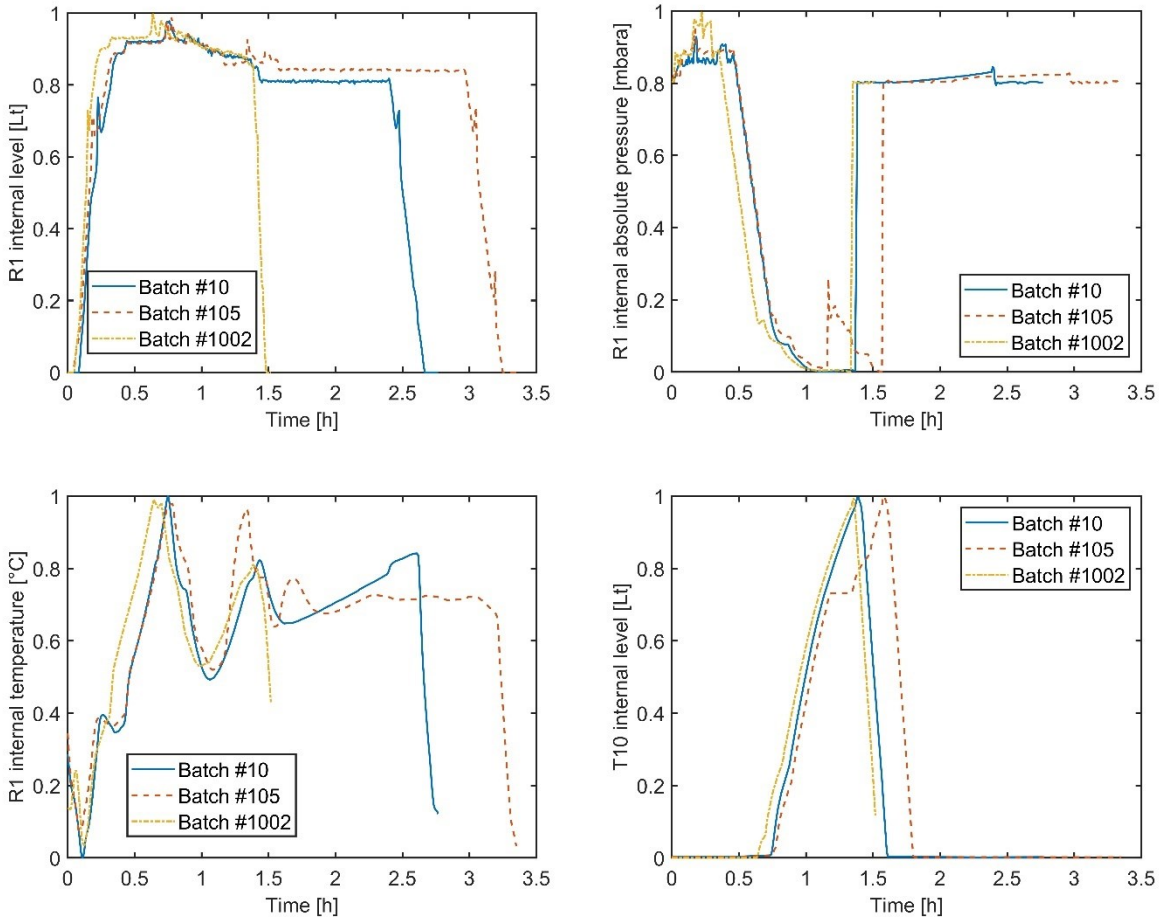


Figure 2.6 *Process variables time trajectories of (a) reactor R1 internal level; (b) reactor R1 internal absolute pressure; (c) reactor R1 internal temperature; (d) tank T10 internal level.*

In Figure 2.6 typical data acquired from the studied process are shown. The units are anonymised for confidentiality reasons. All the PV signals are affected by noise, missing values and outliers, caused by unintended interruptions of the sensor connections and sensor faults. Furthermore, process upsets, such as the intervention of safety interlocks may affect the time trajectories of process variables and render the process less reproducible in terms of batch length (as clearly shown in Figure 2.6a), variations on the time trajectories (as shown in Figure 2.6b

and 2.6d). Furthermore, changes in the control strategy for some variables can bring to differences in the time trajectories of the measured process variables (as shown in Figure 2.6c).

2.4.2 Product quality measurements

The product quality is measured in two different points of the process depicted in Figure 2.2:

- after filtering out the catalyst;
- at the end of the process.

According to process operators jargon, the former is called “filtrate”, while the latter is called “distillate”. A sample of the filtrate is taken from the process at irregular periods of time, alternating periods where they are taken one every 3-4 batches with periods where 60 batches are produced without product quality measurements. The distillate instead is sampled with a much lower frequency, with 2 samples per month on average (the plant produces a batch every 10 hours on average).

The product quality of the product from both sources is monitored according to the same 6 metrics: the measurement of the concentration of the main product and the concentration of 5 different byproducts obtained through an offline GC-MS analysis.

The illustrated methodology has two main disadvantages: *(i)* it is expensive and time consuming and requires trained personnel in the quality control laboratory; *(ii)* due to disadvantage *(i)* it is not possible to monitor all the produced batches, in fact, only a few are monitored for detecting macroscopic and consistent trends deviating from the desired product quality. Furthermore, the adopted methodology measures the product quality only at the end of the batch, not allowing a quality-related monitoring of batches in real time.

Furthermore, a batch is identified by a unique batch number until it reaches T8: at that point 2 batches are joined for the subsequent evaporation steps, the batch number is lost and also the possibility of analyzing the quality of single batches.

2.5 Advantages and challenges in the application of Industry 4.0 approaches for batch process monitoring in specialty chemicals manufacturing

The adoption of Industry 4.0 approaches for batch process monitoring in specialty chemicals manufacturing can bring significant advantages.

The use of data-driven process monitoring techniques can significantly improve efficiency, quality and overall performance of the manufacturing processes through:

1. process insights: it allows operators and decision-makers to monitor process variables, performance metrics and equipment status, enabling prompt identification of issues or deviations, leading to quick responses and corrective actions;

2. improved process efficiency: it can identify areas of inefficiency or bottlenecks in the production process, empowering the manufacturers to optimize process parameters and reduce cycle times, improving resource utilization and the environmental impact of the processes;
3. root cause analysis: it can support root cause analysis, in particular by finding what are the sensor measurements correlated with the occurrence of a particular fault or with the production of off-spec products;
4. enhanced product quality: it can identify the process parameters closely related to product quality and aid in reducing their variability for improving consistency and quality of the final product.

However, the development and application of Industry 4.0 approaches for batch process monitoring is challenging due to the complexity of the studied process and its flexibility, that allows several sources of variability to enter the process, most of which cannot be eliminated:

- variations in the quality of raw materials;
- variations in the status of equipment and utilities;
- the presence of a wide array of parallel units, each one slightly different from the other with respect to geometry, mixing, heat exchange, status and age, control strategy;
- unknown quality of the catalyst and deactivation due to it being recycled back to the process;
- manual variations of controller set points.

Chapter 3

Mathematical methods

This chapter presents the mathematical techniques utilized in this Dissertation. It offers a succinct overview of multivariate statistical techniques, specifically the theoretical formulation of PCA and PLS(-DA). The chapter then proceeds to provide key concepts of multivariate statistical process monitoring, and briefly discusses how multivariate methods are incorporated into monitoring frameworks of batch processes. Finally, the fundamental principles of surrogate optimisation are summarised, and the process for calculating the radial basis function (RBF) interpolator, one of the most frequently used surrogate models, is detailed.

3.1 Multivariate statistical techniques

The mathematical and statistical foundations of the multivariate statistical methods employed in this Dissertation are explored in the next sections. Specifically, we elaborate on the theoretical and algorithmic aspects of principal component analysis (PCA) and projection onto latent structures (PLS).

3.1.1 Principal component analysis

Principal component analysis (PCA; Jackson, 1991) is a multivariate statistical methodology commonly adopted for summarizing the information of a large set of correlated observable variables on a few orthogonal unobservable variables called latent variables.
Let $\mathbf{X}[I \times J]$ be a historical dataset of I observations of J variables conveniently pre-treated (e.g. mean-centered and variance-scaled). The covariance of matrix $\mathbf{X}$ is defined as

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{I-1} \ , \tag{3.1}$$

where $\mathbf{X}^{\mathrm{T}}$is the transposed of $\mathbf{X}$. The PCA method produces an eigendecomposition of the covariance matrix (Wise and Gallagher, 1996):

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_a = \boldsymbol{\lambda}_a\mathbf{p}_a \ , \tag{3.2}$$

where for each latent variable a an eigenvector $\mathbf{p}_a[J \times 1]$ called loadings vector is generated. The loadings vector describes the direction of the a-th latent variable. A scores vector $\mathbf{t}_a[I \times 1]$ is associated to the loadings vector according to:

$$\mathbf{X}\mathbf{p}_a = \mathbf{t}_a \ . \tag{3.3}$$

The loadings vectors are orthogonal; the first one captures the direction of greatest variance of the dataset, the second loadings vector captures the direction of greatest variance orthogonal to the first one and so on (Dunn, 2019). The **X** matrix can be reconstructed as the sum of the outer products of the $(\mathbf{t}_a; \mathbf{p}_a)$ pairs as:

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \ , \tag{3.4}$$

where $A \leq \min(I, J)$ is the number of considered principal components, **T** and **P** are the matrices collecting the scores and the loadings vector, respectively and **E** is the residual matrix, containing the variability of the original matrix not described by the model. The decomposition of the **X** dataset carried out by PCA is shown in Figure 3.1.

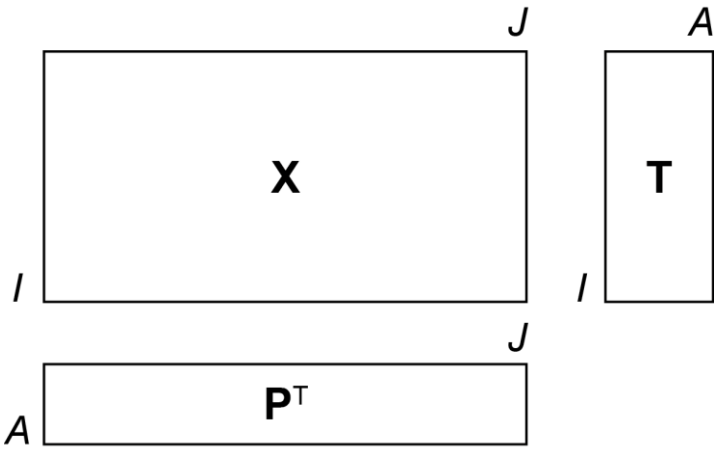


Figure 3.1: *Dataset decomposition carried out by principal component analysis (PCA)*

Formally, the problem solved by PCA for each principal component is (Bro and Smilde, 2014):

$$\underset{\|\mathbf{p}_a\|=1}{\operatorname{argmax}}(\operatorname{var}(\mathbf{t}_a)) \ , \tag{3.5}$$

which consists in the problem of finding the $\mathbf{p}_a$ with unitary Euclidean norm (indicated by the symbol $\|\cdot\|$). Substituting $\mathbf{t}_a$ through (3.3) the problem becomes more explicit:

$$\underset{\|\mathbf{p}_a\|=1}{\operatorname{argmax}}(\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a) = \underset{\|\mathbf{p}_a\|=1}{\operatorname{argmax}}(\mathbf{p}_a^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}_a) \ , \tag{3.6}$$

The problem stated in (3.6) is a standard linear algebra problem where the optimal $\mathbf{p}_a$ is the first eigenvector of the covariance matrix of **X**. Several methods have been proposed to choose the appropriate number of principal components *A* retained by the model (Brown, 2009). Among the most effective methods for the choice are the scree test (Jackson, 1991) and cross-validation (Bro *et al.*, 2008). Refer to Camacho and Ferrer (2014) on several methods to determine the number of principal components based on the objective function to optimize. For

more details the reader is invited to refer to the original references. For each row of $\mathbf{X}$ it is possible to assess a lack of model fit statistic called squared prediction error (Q), considering the sum of squares of the residual matrix, $\mathbf{E}$:

$$Q_i = \mathbf{e_i}\mathbf{e_i}^{\mathrm{T}} \ , \tag{3.7}$$

where $\mathbf{e}_i$ is the i-th row of $\mathbf{E}$. This statistic evaluates the model representativeness, namely how well the model fits the actual conditions of the i-th observation. In other words, Q is a measure of how well each sample conforms to the PCA model (Wise and Gallagher, 1996). It is also possible to calculate a statistic assessing the deviation of the projected i-th observation onto the PCA model from the average condition of the observations in $\mathbf{X}$. To this purpose, for the i-th observation the Hotelling's T^2 statistic is calculated as

$$T_i^2 = \sum_{a=1}^{A} \left(\frac{t_{i,a}}{s_a}\right)^2 \ , \tag{3.8}$$

where s_a is the standard deviation of the a-th scores vector. Once a new observation $\mathbf{x}_{NEW}^{\mathrm{T}}[1 \times J]$ is available, it is projected to the PCA through

$$\mathbf{t}_{NEW}^{\mathrm{T}} = \mathbf{x}_{NEW}^{\mathrm{T}}\mathbf{P} \ , \tag{3.9}$$

Where $\mathbf{t}_{NEW}^{\mathrm{T}}$ is the row vector of scores of the new observation.

3.1.2 Projection onto latent structures

Projection onto latent structures (PLS; Wold *et al.*, 1983), also called partial least-squares regression, is a multivariate statistical technique used for solving regression problems with noisy, multicollinear data (Indahl, 2014). Being a regression methodology, it is used to correlate a predictor dataset $\mathbf{X}$ to a response dataset $\mathbf{Y}[I \times L]$. It is assumed the observed data to be generated by a system or a process with a set of driving forces much smaller than the number of observed variables, hence the number of latent variables (Rosipal and Krämer, 2006):
$A \ll \min(I, J, L)$. In summary, PLS is a methodology explaining the directions of maximum variability of $\mathbf{X}$ that better predict $\mathbf{Y}$. The structure of the PLS model is summarized by the following set of equations:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \ , \tag{3.10}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \ , \tag{3.11}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \ , \tag{3.12}$$

Where $\mathbf{T}[I \times A]$ is the score matrix, $\mathbf{P}[J \times A]$ and $\mathbf{Q}[L \times A]$ are, respectively, the $\mathbf{X}$ and $\mathbf{Y}$ loading matrices and $\mathbf{W}[J \times A]$ is the weight matrix used for projecting the data in $\mathbf{X}$ onto the latent space to calculate $\mathbf{T}$ according to (3.9). The structure of the PLS model is shown in Figure

3.2. The PLS components are extracted sequentially. The calculation of the weight vector $\mathbf{w}_a[J \times 1]$ requires solving the following eigendecomposition problem:

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_a = \lambda_a \mathbf{w}_a \ , \tag{3.13}$$

where λ_a is the eigenvalue associated with the a-th LV (Höskuldsson, 1996), which is equivalent to the following optimization problem (Höskuldsson, 1988):

$$\begin{aligned} &\max \mathbf{w}_a{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w}_a \\ &subject\ to \\ &\mathbf{w}_a{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}_a = 1 \quad . \end{aligned} \tag{3.14}$$

Solving the problem (3.10) is not straightforward from an algebraic point of view. Hence, several iterative algorithms have been proposed in the literature for a computationally efficient implementation of PLS. The most commonly used algorithm is the nonlinear iterative partial least squares (NIPALS) algorithm described in details in Section 3.1.2.1 (Geladi and Kowalski, 1986).

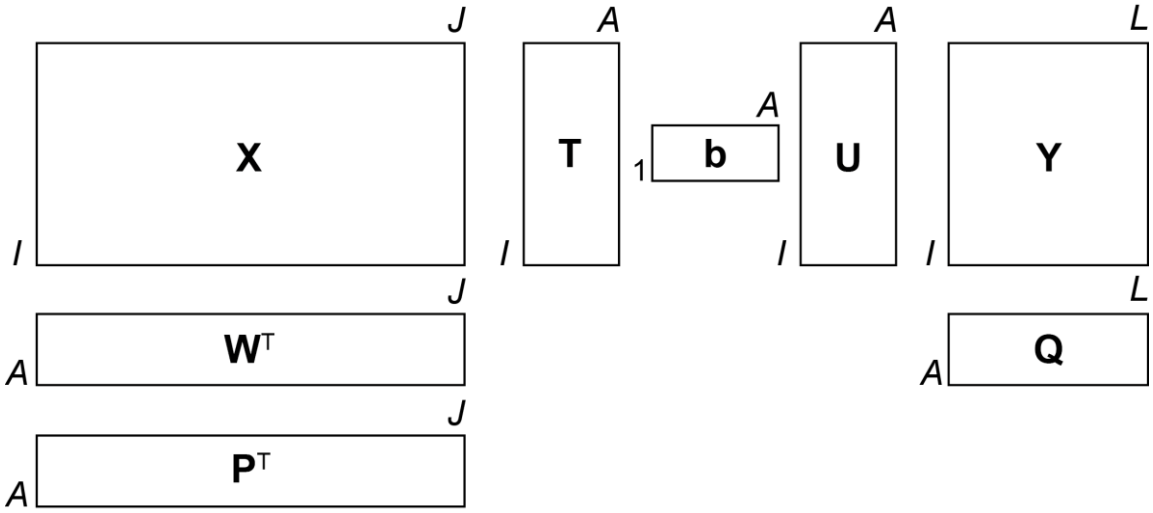


Figure 3.2: *Structure of the Projection onto latent structures model (PLS; adapted from Geladi and Kowalski, 1986)*

It must be noted that the **X** score of (3.11) is sometimes substituted by the **Y** score matrix $\mathbf{U}[I \times A]$. However the two formulations are equivalent as a linear relationship (*inner relation*, described in Section 3.1.2.1) relates the two scores matrices. In PLS models it is possible to calculate the Q and T^2 statistics with (3.7) and (3.8), however, usually, the lack-of-fit statistic is calculated also on the **Y** residuals as

$$= \mathbf{f}_i \mathbf{f}_i^{\mathrm{T}} \ . \tag{3.15}$$

The number of retained LVs is chosen by cross-validation (Wold, 1978). Cross-validation is a procedure according to which a PLS model is iteratively calibrated on a portion of the calibration dataset and it is used to predict the response of the excluded samples. Once the

predictions for all samples in the calibration dataset have been obtained, the root-mean square error of cross validation (RMSECV) is calculated as

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I-1}} \tag{3.16}$$

where y_i is the response for observation i, and $\hat{y}_i$ is the response predicted in cross-validation. Once a new observation $\mathbf{x}_{NEW}^{\mathrm{T}}[1 \times J]$ is available, it is projected in the PLS latent space through

$$\mathbf{t}_{NEW}^{\mathrm{T}} = \mathbf{x}_{NEW}^{\mathrm{T}}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \ , \tag{3.17}$$

where $\mathbf{t}_{NEW}^{\mathrm{T}}[1 \times A]$ is the row vector of scores of the new observation. The predicted response for the new observation is calculated as

$$\hat{\mathbf{y}}_{NEW}^{\mathrm{T}} = \mathbf{t}_{NEW}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}} \ , \tag{3.18}$$

where $\hat{\mathbf{y}}_{NEW}^{\mathrm{T}}[1 \times L]$ is the response predicted by the model for the new observation $\mathbf{x}_{NEW}^{\mathrm{T}}$.

3.1.2.1 NIPALS algorithm

In order to solve the optimization problem (3.11) it is not enough to calibrate a PCA model over **X** and **Y** and build a relation between the obtained principal components. For this reason, the NIPALS algorithm (Geladi and Kowalski, 1986) have been proposed with the following structure, for each component:

1. select a column of $\mathbf{F}_{a-1}(\mathbf{F}_0 = \mathbf{Y})$ as an initial estimate of $\mathbf{u}_a$, $\mathbf{u}_{a,\mathrm{start}}$;
2. regress each column of $\mathbf{E}_{a-1}(\mathbf{E}_0 = \mathbf{X})$ over $\mathbf{u}_a$ to obtain $\mathbf{w}_a$: $\mathbf{w}_a^{\mathrm{T}} = \mathbf{u}_a^{\mathrm{T}}\mathbf{E}_{a-1}/\mathbf{u}_a^{\mathrm{T}}\mathbf{u}_a$;
3. normalize $\mathbf{w}_a$: $\mathbf{w}_{a,\mathrm{new}}^{\mathrm{T}} = \mathbf{w}_{a,\mathrm{old}}^{\mathrm{T}}/\left\|\mathbf{w}_{a,\mathrm{old}}^{\mathrm{T}}\right\|$;
4. regress each row of $\mathbf{E}_{a-1}$ over the weight vector to obtain $\mathbf{t}_a$: $\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{w}_a/\mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a$;
5. regress each column of $\mathbf{F}_{a-1}$ over $\mathbf{t}_a$ to obtain **Y** loadings $\mathbf{q}_a$: $\mathbf{q}_a = \mathbf{t}_a^{\mathrm{T}}\mathbf{F}_{a-1}/\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$;
6. normalize $\mathbf{q}_a$: $\mathbf{q}_{a,\mathrm{new}}^{\mathrm{T}} = \mathbf{q}_{a,\mathrm{old}}^{\mathrm{T}}/\left\|\mathbf{q}_{a,\mathrm{old}}^{\mathrm{T}}\right\|$;
7. regress each row of $\mathbf{F}_{a-1}$ over $\mathbf{q}_a$ to obtain **Y** scores: $\mathbf{u}_a = \mathbf{F}_{a-1}\mathbf{q}_a/\mathbf{q}_a^{\mathrm{T}}\mathbf{q}_a$.

After this step convergence is checked by comparing the obtained $\mathbf{u}_a$ with $\mathbf{u}_{a,\mathrm{start}}$. If $\left\|\mathbf{u}_a - \mathbf{u}_{a,\mathrm{start}}\right\| < \epsilon$, where ϵ is an arbitrary tolerance value (usually set at 10^{-10}) then the successive steps are executed, otherwise the algorithm goes back to step 2:

8. calculate **X** loadings: $\mathbf{p}_a^{\mathrm{T}} = \mathbf{t}_a^{\mathrm{T}}\mathbf{E}_{a-1}/\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$;
9. normalize $\mathbf{p}_a$: $\mathbf{p}_{a,\mathrm{new}}^{\mathrm{T}} = \mathbf{p}_{a,\mathrm{old}}^{\mathrm{T}}/\left\|\mathbf{p}_{a,\mathrm{old}}^{\mathrm{T}}\right\|$;
10. normalize $\mathbf{t}_a$ and $\mathbf{w}_a$ accordingly by multiplying them by $\left\|\mathbf{p}_{a,\mathrm{old}}^{\mathrm{T}}\right\|$;
11. calculate the inner relation coefficient: $b_a = \mathbf{u}_a^{\mathrm{T}}\mathbf{t}_a/\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$.

In the last two steps the calculated PLS component is removed from the residual matrices obtained for the previous component in order to obtain the new residual matrices:

12. $\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$;
13. $\mathbf{F}_a = \mathbf{F}_{a-1} - b_a\mathbf{t}_a\mathbf{q}_a^{\mathrm{T}}$.

These matrices are used to calculate the elements of the subsequent latent variable repeating the algorithm iteratively.

3.1.3 Projection onto latent structures discriminant analysis

The extension of PLS to classification problems is the so-called projection onto latent structures discriminant analysis (PLS-DA; Barker and Rayens, 2003). Consider a response dataset $\mathbf{Y}_c[I \times L]$ where the element in row i and column l is equal to 1 if the i-th observation belongs to the l-th class (e.g., an on-spec product), or 0 otherwise (e.g., an off-spec product). PLS-DA works by building a PLS model on datasets $\mathbf{X}$ and $\mathbf{Y}_c$. The estimated class attribution obtained in calibration, $\widehat{\mathbf{Y}}_c$, are used to fit a cumulative density function to identify the probability of belonging to a specific class. Once the prediction on a new observation is calculated through (3.14), the abovementioned cumulated density functions are used to calculate the probability of attributing the new observation to the relevant class (Ballabio and Consonni, 2013).

3.1.4 Process monitoring with multivariate statistical techniques

MSPM carries out the activities constituting process monitoring described in Section 1.4 through the use of a model describing the process NOC developed with multivariate statistical modelling techniques (MacGregor and Kourti, 1995). A dataset of historical NOC observations composed of plant sensors measurements $x_1, \dots x_J$ is retrieved from the plant data historian. Under NOC, the attributes of a process are assumed to stay close to a assigned target value without changing perceptibly. MSPM assumes therefore that the process adheres to a state of statistical control if it remains within certain statistical limits. Once a multivariate statistical model is calibrated on historical data, confidence limits can be established for T^2 and Q. In particular, the one-sided Q confidence interval is calculated as follows (Eriksson *et al.*, 2006)

$$Q_{lim} = \frac{\sigma}{2\mu} \chi^2_{\frac{2\mu}{\sigma},\alpha} \quad , \tag{3.19}$$

where μ and σ are the mean and the variance of the residuals from the calibration dataset, $\mathbf{X}$, and χ^2 is the chi-squared distribution with $2\mu/\sigma$ degrees of freedom and significance α. The one-sided confidence interval for the Hotelling's T^2 statistic is calculated as follows (Wise and Gallagher, 1996):

$$T^2_{lim} = \frac{A(N-1)}{(N-A)} \mathrm{F}_{A,N-A,\alpha} \quad , \tag{3.20}$$

where F is a Fisher distribution with A numerator degrees of freedom, $N - A$ denominator degrees of freedom and significance α. When a new observation is available, its Q statistic value is calculated as

$$Q_{NEW} = \mathbf{e}^{\mathrm{T}}_{NEW} \mathbf{e}_{NEW}, \tag{3.21}$$

where Q_{NEW} is the Q_{NEW} statistic value of the new observation and $\mathbf{e}_{NEW}$ is the vector of the residuals of the new observation. Its T^2 statistic is calculated as follows:

$$T^2_{NEW} = \sum_{a=1}^{A} \left(\frac{t_{NEW,a}}{s_a} \right)^2 , \tag{3.22}$$

where $t_{NEW,a}$ is the value of the score of the new observation on the a-th latent variable and s_a is the a-th latent variable calibration scores variance. For fault detection purposes, Q_{NEW} and T^2_{NEW} values are compared with Q_{lim} and T^2_{lim} values, respectively. If either of the two statistics are out of the relevant confidence limit, the observation is assumed to be faulty. A geometrical interpretation of the SPE and T^2 confidence limits as used for MSPM, hence through a PCA or PLS model, is given in Figure 3.3, where the observations from a reference dataset are projected from the original space onto a latent space, where the latent variables are the directions of maximum variability of the data. Within this sub-space of the original space, the compliance of new observations can be analyzed through T^2 and Q statistics.

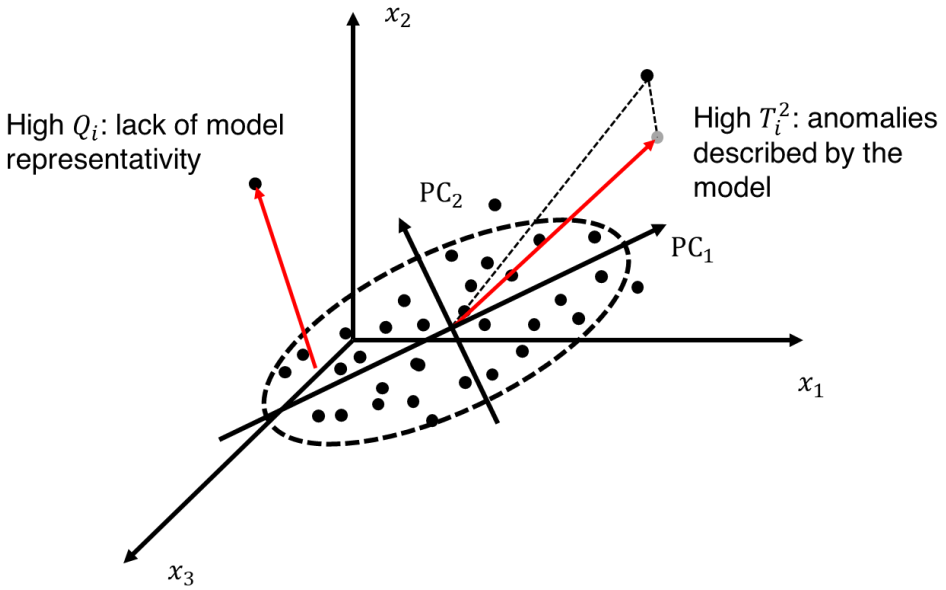


Figure 3.3: *Geometrical interpretation of the T^2 and Q statistics in multivariate statistical process monitoring (MSPM)*

In particular, the T^2 statistic is a measure of the distance of the projection of a given observation onto the PCA model from the average condition of the reference (the origin of the latent variables axes), while the Q statistic indicates the distance of the new observation from the latent variables hyperplane.

3.1.4.1 Contribution plots

While carrying out fault detection in MSPM, the comparison of statistic values with their calculated confidence limits during multivariate statistical model calibration is undertaken.

Fault diagnosis and identification is carried out assessing the correlations between the original J variables in dataset $\mathbf{X}$ and the occurrence of the observed anomalous condition. This task can be aided by the calculation of the contributions to the observed value of the T^2 and Q statistic, as they help to make a sound guess for the assignable causes of the observed anomaly (Nomikos, 1996). Contribution plots (Miller *et al.*, 1998) are the most commonly adopted approach for detecting the root cause of the fault. The calculation of the Q contribution is straightforward, as they are simply the elements of the residuals vector $\mathbf{e}_{NEW}$. The contributions to T^2_{NEW} are obtained as follows:

$$\mathbf{t}^2_{\mathrm{con,NEW}} = \mathbf{x}^{\mathrm{T}}_{\mathrm{NEW}}\mathbf{P}\mathbf{S}^{-\frac{1}{2}}\mathbf{P}^{\mathrm{T}} \quad , \tag{3.23}$$

where $\mathbf{t}^2_{\mathrm{con,NEW}}[1 \times J]$ is the vector of the T^2_{NEW} contributions, $\mathbf{x}^{\mathrm{T}}_{\mathrm{NEW}}[1 \times J]$ is the new observation, $\mathbf{S}[J \times J]$ is a matrix containing the variance of the columns of $\mathbf{T}$ on its main diagonal. The variables with the larger absolute values of the contributions relative to the statistic that violated its control limit are considered correlated with the fault occurrence (Joe Qin, 2003).

3.1.5 Batch process monitoring with multivariate statistical techniques

When monitoring batch processes, the nature of batch data must be understood first. In a batch process, the operating conditions are time-dependent, hence, a third dimension must be considered in batch data, namely, time. Let $\underline{\mathbf{X}}_{\mathrm{B}}[I \times J \times \widetilde{K}]$ be a three-dimensional tensor of historical data from$I$ batches. For each batch, J process variables are measured at $\widetilde{K}$ time points. Often in real industrial processes each batch have a different time duration, hence, in general, $\widetilde{K}$ changes across batches. In order to perform process monitoring with such data framework using multivariate statistical methods, such as PCA or PLS(-DA), $\underline{\mathbf{X}}_{\mathrm{B}}$ must be transformed into a two-dimensional matrix, and this can be accomplished by resorting to multiway PCA (MPCA) and PLS (MPLS; Nomikos and MacGregor, 1994, 1995a). MPCA and MPLS are consistent with their counterparts both from an algorithmic and a mathematical standpoint. These methods unfold $\underline{\mathbf{X}}_{\mathrm{B}}$ into a two-dimensional matrix that can be processed by standard multivariate techniques.

There are two main ways of unfolding $\underline{\mathbf{X}}_{\mathrm{B}}$:

- batch-wise unfolding (BWU);
- variable-wise unfolding (VWU).

The BWU unfolding slices the batch data tensor along the time direction, obtaining $[I \times J]$ matrices that are placed side-by-side (Nomikos and MacGregor, 1995a).

As observed in Section 1.4.1.2, BWU is the most logical method for modelling differences among batches, especially in the common situation where a single quality measurement at the end of the batch is available for each batch (Golshan *et al.*, 2010). Unfortunately, several drawbacks affects the application of BWU when using it for batch MSPM. In order to use it,

the batch data must be aligned, namely, they must be equalized (i.e. all variables must be available at the same sampling rate) and synchronized (i.e. all the variables trajectories must have the same number of samples across all batches; González-Martínez *et al.*, 2018), this problem will be addressed in details in Section 3.1.5.1.

Furthermore, for online monitoring applications, BWU unfolding cannot be directly applied as it requires data for the entire batch to be available, and this occurs only after the completion of the batch itself. The latter problem can be solved by filling the incomplete matrix for the future unknown samples, at the conditions of having at least 10% of the batch history already available (Nomikos and MacGregor, 1995a).

One of the workarounds for reducing the complexity of the preprocessing needed for monitoring batch processes through multiway techniques is unfolding the 3-dimensional tensor of batch data in the variable direction, thus obtaining the $\mathbf{X_V}[I\widetilde{K} \times J]$ VWU matrix. Thus, in VWU, each batch operating conditions are represented by a time trajectory in the latent (score) space (Wold *et al.*, 1998). However, it is well known that the VWU based multiway models exhibit worse monitoring performance than their BWU counterparts (Kourti, 2003). In order to improve their performance, when using the VWU methodology, in commercial applications it is suggested to build a control chart by building control limits around the time trajectories of NOC calibration scores plotted against sample number (that is equivalent to plotting them against a time axis) for each LV (Eriksson, 2021). The underlying assumption of this methodology is that scores from a multivariate model can be used for building a univariate control chart: this assumption does not always hold and exposes the practitioner to the same risks observed when applying classical univariate methodologies to inherently multivariate processes (MacGregor and Kourti, 1995). Furthermore, both VWU and BWU multiway techniques assume that time (intended in absolute terms) is an attribute of a batch trajectory.

In order to overcome on the one hand the necessity of batch alignment and phase partition of BWU, and on the other hand the unsatisfactory monitoring performance of VWU, an assumption free methodology has been proposed (Westad *et al.*, 2015) and will be discussed in Section 3.1.5.3.

3.1.5.1 Batch alignment

Batch alignment can be challenging. Effective methodologies are necessary for synchronizing batch data for developing MPCA or MPLS(-DA) models through BWU. Ad-hoc synchronization techniques, such as truncating the trajectories of all batches to the shortest batch length (Rothwell *et al.*, 1998) or extending the length of shorter batches by repeating the last measurement (Lakshminarayanan *et al.*, 1996) are simple workarounds that can be set up quickly for preliminary dataset screening and analysis, but may provide ineffective data modelling (Rendall *et al.*, 2019). More advanced methodologies are based upon techniques borrowed from speech recognition, such as dynamic time warping (DTW; Kassidas *et al.*, 1998)

and its offspring both for online (relaxed greedy optimized warping, RGTW; González-Martínez *et al.*, 2011) and offline (multisynchro, MS; González-Martínez *et al.*, 2014) synchronization, or from spectroscopy, such as correlation optimized warping (COW; Fransson and Folestad, 2006; Tomasi *et al.*, 2004). These advanced methodology, although being highly effective, have well-known downsides, such as a high computational cost with large datasets (Mueen and Keogh, 2016) and the generation of artifacts when some batches are significantly shorter than the chosen reference (José M. González-Martínez *et al.*, 2014) and/or the starting and ending process conditions are different (González-Martínez *et al.*, 2011). All the previously mentioned methodologies require some process knowledge to be applied, e.g. choosing a reference batch, a reference variable, or both. This can be problematic in processes where first principles knowledge is scarce, such as batch processes for the production of specialty chemicals.

A more effective, yet still simple, synchronization strategy consists in nonlinearly mapping time to an indicator variable (IV; Nomikos and MacGregor, 1995b). In this approach, a variable is selected for representing the "maturity" of a batch and batches are aligned according to the percentage of final value of this variable attained at the current point in time.

The indicator variable methodology is the most popular technique due to its ease of implementation and effectiveness (Barton *et al.*, 2021; Brunner *et al.*, 2020; Kourti *et al.*, 1996; Krause *et al.*, 2015; Neogi and Schlags, 1998). A process variable must possess three properties in order to be used as an indicator variable: *i*) it must be monotonic (García-Muñoz *et al.*, 2003); *ii*) it must have a sufficiently high signal-to-noise ratio (Ündey *et al.*, 2003); *iii*) it must have approximately the same initial and final values across all batches in $\underline{\mathbf{X}}_{\mathrm{B}}$
(Nomikos and MacGregor, 1994). The appropriate variable is selected using process knowledge (García-Muñoz *et al.*, 2011; Kourti, 2003). A single IV may not exist for an entire batch, but can exist for single time windows wherein the measured variables have similar correlation structure (i.e. a different IV may exist in each batch process phase), for this reason, a phase partitioning algorithm must be applied before aligning batches with the IV technique.

3.1.5.2 Phase partition

Also partitioning a batch into phases is a challenging task (Camacho *et al.*, 2008a; Guo and Jin, 2019; Lu *et al.*, 2004; Luo *et al.*, 2016; Zhang *et al.*, 2018), because phases do not always match the physical events in a process (i.e., phases do not always correspond to processing steps).

Guo and Jin (2019) proposed a model-agnostic, phase partition methodology that automatically returns the number Φ_i of phases into which one entire non-synchronized batch i can be partitioned, together with the time point at which each phase onsets. The methodology, schematically shown in Figure 3.4, is based on analyzing the change in the correlation structure of the measured data across multiple consecutive time points. Here, we summarize the original methodology proposed by Guo and Jin (2019).

Consider a horizontal slice of $\underline{\mathbf{X}}_{\mathrm{B}}$, including all measurements taken from batch i across all K_i time points along the batch, and arrange the relevant data in matrix $\mathbf{X}_i[J \times K_i]$. A moving window $\mathbf{X}_{i,k}$ $[J \times V]$ of data in $\mathbf{X}_i$ (V being the moving window width) is slid along time, one time point at a time across all measurements, from $k = 1$ up to $k = K_i - V + 1$, so that each time point is included in at least one of the windows. Consider the correlation matrix $\mathbf{C}_k$ $[J \times J]$ of $\mathbf{X}_{i,k}$; its generic (p, q) element is calculated as:

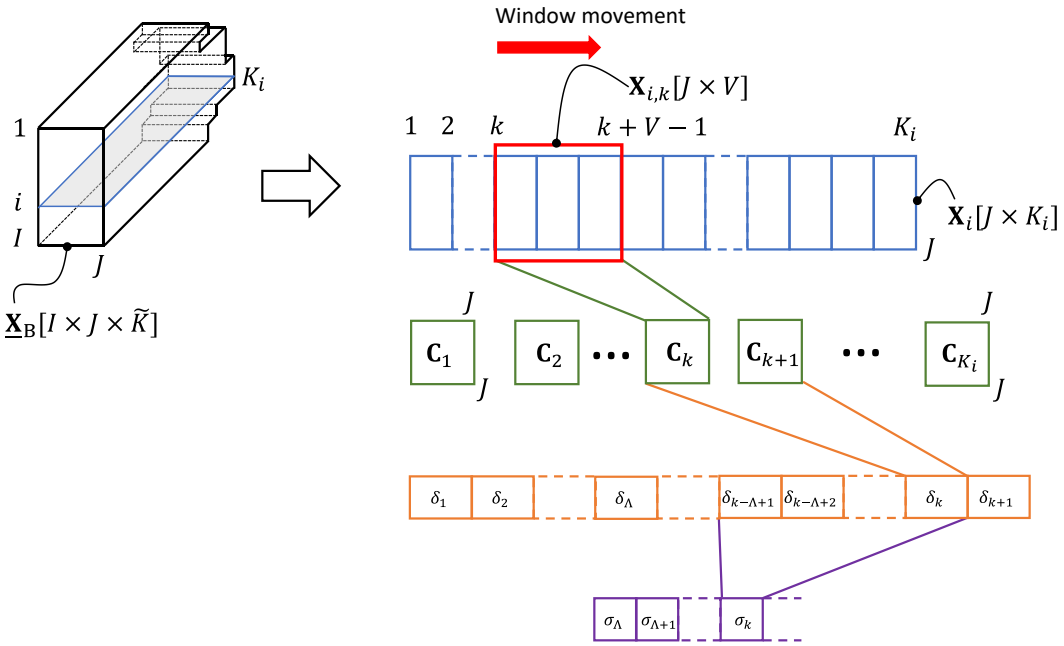


Figure 3.4. *Procedure performed for automatic phase partition of multiphase, uneven-length batches (adapted from Guo and Jin, 2019)*

$$\mathbf{C}_k(p,q) = \frac{\mathrm{cov}\left(\mathbf{X}_{i,k}^p, \mathbf{X}_{i,k}^q\right)}{\sigma(\mathbf{X}_{i,k}^p)\sigma(\mathbf{X}_{i,k}^q)}, \tag{3.24}$$

where $\mathbf{X}_{i,k}^p$ and $\mathbf{X}_{i,k}^q$ are the p-th and the q-th rows in $\mathbf{X}_{i,k}$ (respectively), $\mathrm{cov}\left(\mathbf{X}_{i,k}^p, \mathbf{X}_{i,k}^q\right)$ is the covariance between the previously mentioned rows, and $\sigma(\mathbf{X}_{i,k}^p)$ and $\sigma(\mathbf{X}_{i,k}^q)$ are the standard deviations of the p-th and of the q-th rows in $\mathbf{X}_{i,k}$, respectively.

Define the multidimensional average gain index δ_k between two consecutive correlation matrices as:

$$\delta_k = \frac{\sum_{p=1}^{J}\sum_{q=1}^{J}|\mathbf{C}_{k+1}(p,q) - \mathbf{C}_k(p,q)|}{J^2}. \tag{3.25}$$

The gain captures the variation of the process characteristics (i.e., the change in correlation structure) between consecutive time points as the batch time progresses. A necessary condition to be fulfilled at time point k to trigger the switch from the current phase to a new one is that Λ consecutive values of δ_k exceeds a threshold value Θ_k.

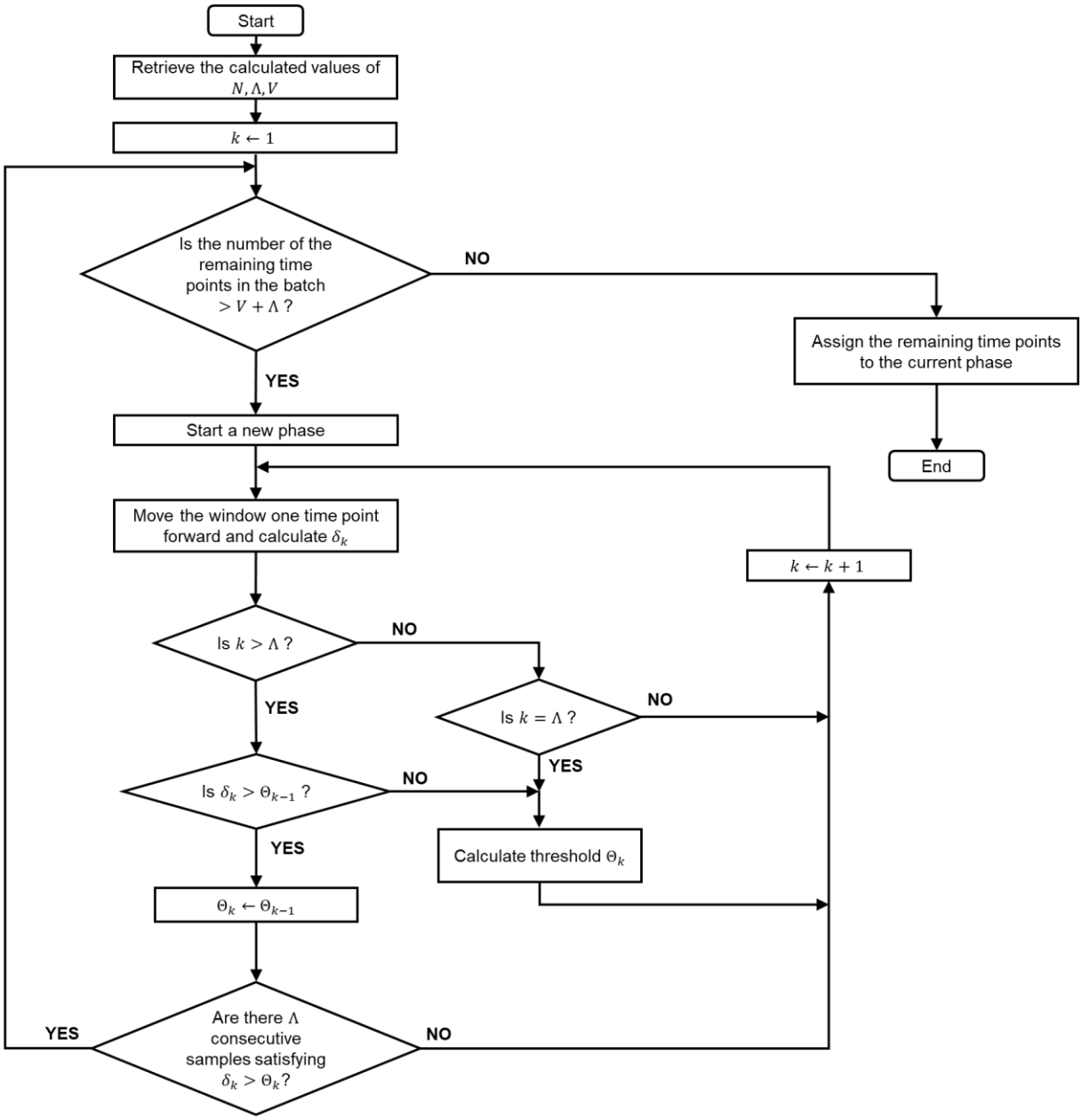


Figure 3.5. *Flow chart of the original automatic phase partition methodology for one entire batch (adapted from Guo and Jin, 2019).*

If the last calculated value of δ_k does not exceed the threshold Θ_{k-1} calculated at the previous window slide, the threshold is calculated from the switch control limit σ_k, defined for a given phase as:

$$\sigma_k = \frac{1}{\Lambda} \sum_{z=k-\Lambda+1}^{k} \delta_z \quad , \tag{3.26}$$

where Λ is the number of time points over which δ_k is averaged. The threshold Θ_k is calculated as:

$$\Theta_k = N\sigma_k \quad , \tag{3.27}$$

where N is a parameter called tolerance factor. Otherwise, if $\delta_k > \Theta_{k-1}$, the threshold Θ_k takes the same value as Θ_{k-1}.

To make phase switch actually occur, condition $\delta_k > \Theta_k$ must be satisfied for Λ consecutive time points. It can be shown that the minimum length of a phase that can be detected by this method is $(V + \Lambda)$ time points. A flowchart of the phase partition mechanism for a generic batch is shown in Figure 3.5.For a given set of batches, the method requires assigning three adjustable parameters, namely the moving window width V, the width Λ over which the gains are averaged for each phase, and the tolerance factor N. The achieved phase partition strongly depends on the values assigned to the parameters; therefore, their search is best done by optimization.

3.1.5.3 Assumption-free monitoring

The assumption-free monitoring technique is a methodology that can accommodate uneven batch lengths, unknown initial batch (absolute) time, phase changes and uneven residence time (Westad *et al.*, 2015). The methodology can achieve this by inherently estimating the so-called "relative time", i.e., a measurement of the progress of the underlying chemical, biological and physical phenomena along the process. The methodology is not based on BWU, in fact its first step consists in carrying out a VWU of the original data matrix with column-wise centering and scaling. According to VWU, each score represents a time point of a batch, hence in the score plot (qualitatively represented in Figure 3.6) of this VWU technique each point represents a time point of a batch, and not a whole batch as is the case in BWU-based techniques. The estimation is carried out by gridding a portion of the score space containing the calibration scores and building a common batch trajectory by calculating the mean of the scores contained in each cell of the grid. For monitoring purposes, a dynamic control limit is calculated around the common batch trajectory. A qualitative representation of the methodology is given in Figure 3.6. The assumption-free monitoring technique is analogous to previously described batch process monitoring techniques as it requires a set of NOC batches to calibrate a VWU MPCA model, whose scores over the first 2 PCs are represented as green upward triangles, yellow circles and orange downward triangles. The grey dashed grid is then built in the latent (score) space of the PCA model and the valid cells of the grid (in light green) are identified.

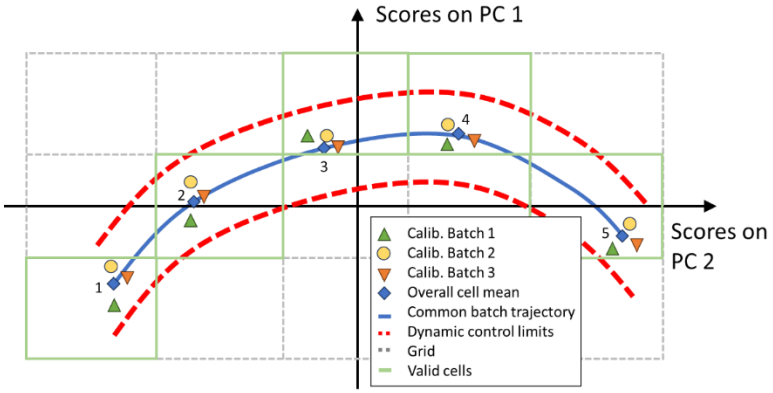


Figure 3.6. *Qualitative representation of the assumption free monitoring methodology*

The mean value of the scores in the valid cells (blue diamonds) are used to build the common batch trajectory (blue line) and to estimate the relative time. The distance of the scores from the common batch trajectory are used to build control limits around the common batch trajectory (dynamic control limits, dashed red lines).

For monitoring purposes, each observation belonging to a new batch are projected onto the PCA model hyperplane and are compared with the control limits.

The detailed algorithm for calibrating the assumption-free monitoring methodology is given as shown in Figure 3.7a and 3.7b:

A1.1.the calibration data tensor $\underline{\mathbf{X}}_{\mathbf{B}}$ is unfolded variable-wise, obtaining $\mathbf{X}_{\mathbf{V}}[I\widetilde{K} \times J]$;

A1.2.$\mathbf{X}_{\mathbf{V}}$ is centered and scaled;

A1.3.a PCA model is calibrated using $\mathbf{X}_{\mathbf{V}}$;

A1.4.a grid is built in the multivariate space for modelling the batch trajectory in the best way.

A1.5.in each cell of the grid, the mean for all observation and the means for individual batches is calculated;

A1.6.the common batch trajectory is built interpolating the overall means of the samples;

A1.7.the individual means in each cell are projected onto the common batch trajectory and relative time, distance within model space and residual distance are estimated;

A1.8.standard deviation around the common batch trajectory is estimated and dynamic control limits are calculated around the common batch trajectory;

A1.9.residual distance and its confidence limit is calculated;

A1.10.relative time for each observation in $\mathbf{X}_{\mathbf{V}}$ is calculated.

With step 10 the model calibration is concluded. When a new observation is available the following steps are followed:

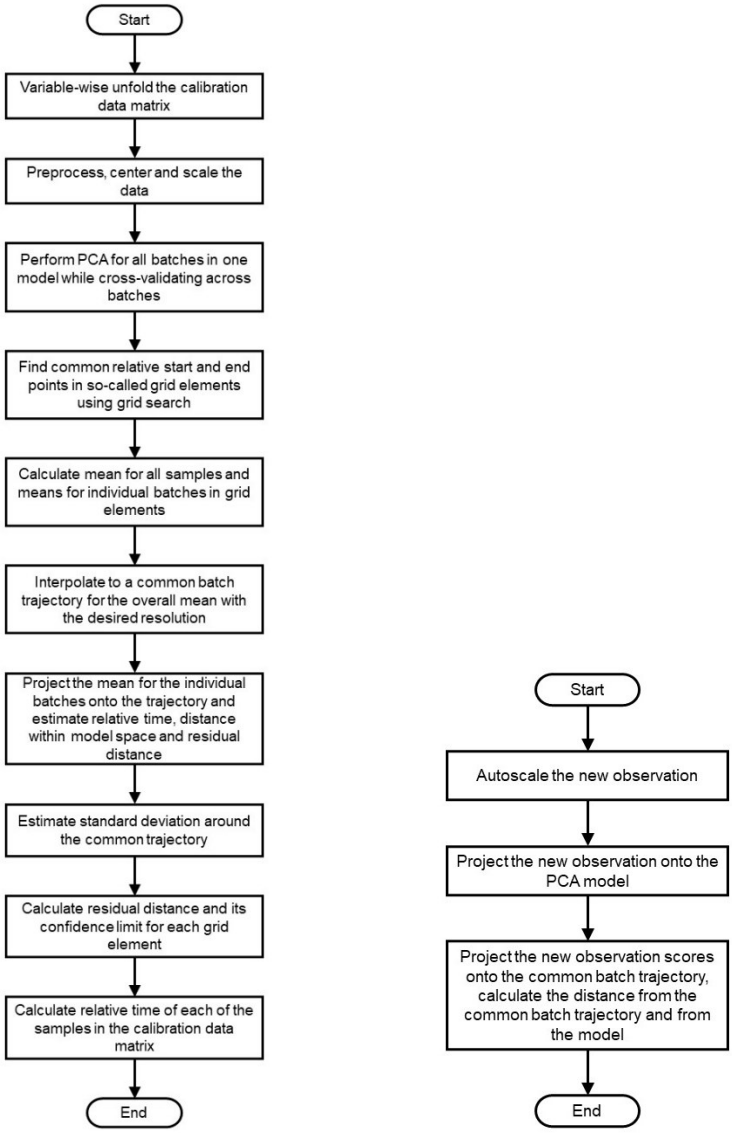
A1.11.the new observation $\mathbf{x}_{\mathbf{V},\mathrm{NEW}}^{\mathrm{T}}[1 \times J]$ is autoscaled;

A1.12.the new score is projected onto the PCA model through (2);

A1.13.the score is projected onto the common batch trajectory calculated in Step 6, its distance from the common batch trajectory and from the model is estimated.

Although the assumption-free monitoring methodology is explained in detail in the original work by Westad *et al.* (2015), some of the steps of the assumption-free monitoring algorithm need further clarifications.

In particular, A1.4 is a key step of the calibration portion of the algorithm: in this Step the "best" grid in the PCA score space is stated to be found through a grid search algorithm, however, the optimization problem that the grid search solves is not formulated explicitly.



(*a*) (*b*)

Figure 3.7: *(a) Flowsheet of steps 1-10 of the Assumption-free monitoring algorithm (calibration). (b) Flowsheet of steps 11-13 of the assumption-free monitoring algorithm (monitoring).*

In Step A1.6 it is required to interpolate the overall means in the valid cells in order to build a common batch trajectory. The original work loosely states that either a linear or a spline interpolant can be used for this task based upon the nature of the studied process, however, a

clear guideline for choosing the interpolant function that achieve the best monitoring results is missing. A clear description on how to perform fault detection and fault identification and diagnosis is completely missing in the original work, although it is implicitly stated that the commercial implementation of the assumption-free monitoring methodology performs fault detection using the dynamic control limits and the residual distance control limits calculated in steps A1.8 and A1.9 and the conventional T^2 contribution plots are used for fault diagnosis (Westad, 2020).

3.2 Surrogate optimization

Surrogate optimization is a global optimization methodology especially useful when the objective functions are non-smooth (Queipo *et al.*, 2005), as occurs for example when the optimization variables are discrete. One significant advantage of surrogate optimization is that it can be applied with unknown symbolic form of the objective function and unknown exact derivatives of the function itself (Bhosekar and Ierapetritou, 2018).

A surrogate $\hat{\mathcal{L}}(\boldsymbol{\xi})$ is obtained that approximates the loss function, has a known analytical form, and is cheaper to evaluate with respect to the true objective function $\mathcal{L}(\boldsymbol{\xi})$. In this study we use the radial basis function (RBF) interpolator, which has the form

$$\hat{\mathcal{L}}(\boldsymbol{\xi}) = \sum_{h=1}^{H} \lambda_h \Psi\left(\left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_h\right\|_2\right) + \wp(\boldsymbol{\xi}) \ , \tag{3.28}$$

where H is the number of parameter sets for which the value of $\mathcal{L}$ is known and upon which the interpolation is made, $\bar{\boldsymbol{\xi}}_h$ is one such parameter set, the λ_i's are weights to be determined by calibration, $\Psi(\cdot)$ is the RBF, $\wp$ is a polynomial whose coefficients are to be determined, and $\|\cdot\|_2$ is the Euclidean norm (Chen *et al.*, 2022). The RBF chosen in this study is the cubic one, which has been proven to outperform other surrogate models (Bano *et al.*, 2018), while the polynomial $\wp$ has degree 1. This RBF has also been proven to minimize a measure of bumpiness (Gutmann, 2001). The optimization algorithm alternates between two phases: surrogate construction, and minimum search.

The surrogate construction consists of these steps:

1. A_1 quasirandom input vectors (i.e., parameter sets) are sampled within the bounds.
2. $\mathcal{L}$ is evaluated on these points. The minimum value of the objective function among these points is identified as the "incumbent".
3. $\hat{\mathcal{L}}$ is calibrated on the values of $\mathcal{L}$ obtained at point 2.

After these steps the minimum search starts:

4. A_2 input vectors are sampled in the input space close to the incumbent.
5. $\hat{\mathcal{L}}$ is evaluated on the points identified in step 4.

6. The point with minimum $\hat{\mathcal{L}}$ in step 5 is identified. This point is added to the initial points of step 1, and the algorithm iterates back to step 2, until an assigned number of iterations is reached.

When the minimum search problem involves both real and integer variables the optimization step must be adjusted to account for this. The most commonly adopted adaptation is a variant of the branch-and-bound mixed-integer optimization algorithm proposed by Achterberg *et al.* (2005).

3.2.1 Radial basis function interpolators

In this Section, we provide a short overview of radial basis functions (RBFs), more details can be found in specialized references (Biancolini, 2017; Iske, 2002).

RBFs have found applications in several domains, such as computer graphics (Zhong *et al.*, 2019), predictive maintenance (She *et al.*, 2020), and chemometrics (Xu *et al.*, 2006), most frequently for scattered data interpolation. The data interpolation problem can be stated as follows: given H multidimensional data points $\bar{\boldsymbol{\xi}}_h$ (with $h = 1, 2, \cdots, H$), with corresponding scalar values $\mathcal{L}(\bar{\boldsymbol{\xi}}_h)$, compute a function $\hat{\mathcal{L}}(\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a generic point belonging to the same space to which the data points $\bar{\boldsymbol{\xi}}_h$ belong, that smoothly interpolates the data points, and for which $\mathcal{L}(\bar{\boldsymbol{\xi}}_h) = \hat{\mathcal{L}}(\bar{\boldsymbol{\xi}}_h)$ for all the values of h.

In order to carry out this task, a function $\Psi(\cdot)$ of the distance between $\bar{\boldsymbol{\xi}}_h$ and $\boldsymbol{\xi}$, called RBF, is used for generalizing the concept that the closer we get to a certain data point $\bar{\boldsymbol{\xi}}_h$, the closer the value of $\hat{\mathcal{L}}$ should get to $\mathcal{L}(\bar{\boldsymbol{\xi}}_h)$. A cubic RBF Ψ is defined as:

$$\Psi(\left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_h\right\|_2) = \left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}_h\right\|_2^{\,3} \quad , \tag{3.29}$$

where $\|\cdot\|_2$ is the Euclidean norm. Thus, the RBF-based interpolator takes the form:

$$\hat{\mathcal{L}}(\boldsymbol{\xi}) = \sum_{h=1}^{H} \lambda_h \Psi(\cdot) \quad . \tag{3.30}$$

Solving this equation consists in solving the following linear system:

$$\begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} & \cdots & \Psi_{1,H} \\ \Psi_{2,1} & \Psi_{2,2} & \cdots & \Psi_{2,H} \\ \vdots & \vdots & \ddots & \\ \Psi_{H,1} & \cdots & & \Psi_{H,H} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_H \end{pmatrix} = \begin{pmatrix} \mathcal{L}(\bar{\boldsymbol{\xi}}_1) \\ \mathcal{L}(\bar{\boldsymbol{\xi}}_2) \\ \vdots \\ \mathcal{L}(\bar{\boldsymbol{\xi}}_H) \end{pmatrix} \Rightarrow \boldsymbol{\Psi\lambda} = \boldsymbol{\mathcal{L}} \tag{3.31}$$

where the element in position (i, j) of matrix $\boldsymbol{\Psi}$ is $\Psi_{i,j} = \Psi(\left\|\bar{\boldsymbol{\xi}}_i - \bar{\boldsymbol{\xi}}_j\right\|_2)$, and the unknowns are the weights λ_h's.

One disadvantage of the RBF interpolator in (3.23) is that it is unable to represent polynomial functions. In order to make it able to approximate polynomial functions, a polynomial function $\wp(\boldsymbol{\xi})$ is appended to the right-hand side of (3.23). For example, a linear polynomial can be used:

$$\wp(\boldsymbol{\xi}) = \mathrm{c}_1 + c_2\boldsymbol{\xi} \quad , \tag{3.32}$$

where c_1 and c_2 are the parameters of the polynomial. Thus, we obtain the final form of the RBF-based interpolator:

$$\hat{\mathcal{L}}(\boldsymbol{\xi}) = \sum_{h=1}^{H} \lambda_i \Psi + \wp(\boldsymbol{\xi}) \quad . \tag{3.33}$$

Let $\boldsymbol{\mathcal{B}}$ be the basis of $\wp$:

$$\boldsymbol{\mathcal{B}} = \begin{pmatrix} 1 & \bar{\boldsymbol{\xi}}_1 \\ 1 & \bar{\boldsymbol{\xi}}_2 \\ \vdots & \vdots \\ 1 & \bar{\boldsymbol{\xi}}_H \end{pmatrix} \quad . \tag{3.34}$$

The linear system to be solved becomes:

$$\boldsymbol{\Psi\lambda} + \boldsymbol{\mathcal{B}}\mathbf{c} = \boldsymbol{\mathcal{L}} \quad , \tag{3.35}$$

where $\mathbf{c} = (c_1\ c_2)^{\mathrm{T}}$. However, the system is now underdetermined. In order to be able to solve (3.28), we constrain the weights $\boldsymbol{\lambda}$ to be zero if the polynomial terms match the data points exactly with the coefficients $\mathbf{d}$:

$$\boldsymbol{\Psi\lambda} + \boldsymbol{\mathcal{B}}\mathbf{c} = \boldsymbol{\mathcal{B}}\mathbf{d} \quad . \tag{3.36}$$

After a few algebraic manipulations, we end up with the following linear system (Biancolini, 2017):

$$\begin{pmatrix} \boldsymbol{\Psi} & \boldsymbol{\mathcal{B}} \\ \boldsymbol{\mathcal{B}}^{\mathrm{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mathcal{L}} \\ \mathbf{0} \end{pmatrix} \quad , \tag{3.37}$$

which can easily be solved by linear algebra.

Chapter 4

Assessment of the benefits of the application of batch process monitoring techniques*

In this chapter, the effectiveness of multivariate batch process monitoring techniques in an industrial setting is assessed. A state-of-the-art methodology for batch process troubleshooting is applied to an industrial batch process manufacturing an intermediate for the production of polymer additives with unresolved issues. It is assessed that the use of multivariate statistical methods on an industrial plant is an effective approach for troubleshooting processes and allow improving the process efficiency and safety.

4.1 Introduction

Batch processes are widespread in many industries producing low volumes of high added-value goods such as pharmaceuticals, biotechnological products and specialty chemicals. The energy costs for common chemicals produced in batch operations can arrive at as high as 10% of the total production costs (Bieler *et al.*, 2004). Therefore, reducing energy consumption not only reduces the environmental impact of the process, but can also reduce significantly the process operating expenses.

One approach for modeling batch processes is based on first-principles models, requiring detailed knowledge about the phenomena occurring in the process. The development of these models is usually expensive, time consuming, hence often prohibitive in an industrial setting. On the other hand, the increased availability of data in the process industry, propelled by the development of sensors and networking technology together with the reduction of the costs of computing equipment, allowed the development of data-driven models for tasks traditionally carried out through knowledge-driven models, thus sensibly reducing time and costs for model development. Batch processing is highly impacted by this approach, especially when the process chemistry is not completely understood, which renders the development of a first-principles model a hard (or even impossible) task.

* Sartori F., Zuecco F., Facco, P., Bezzo F., Barolo M. (2022), Data Analytics Can Help Reduce Energy Consumption in the Industrial Manufacturing of Specialty Chemicals. *Chem. Eng. Trans.* **96**, 229-234

In order to extract process-relevant information from the massive amount of data produced by a modern chemical process, effective data analytics techniques can be used. Multivariate statistical methods, such as principal component analysis (PCA; Jollife and Cadima, 2016) and its multiway extension (Nomikos and MacGregor, 1995a) are extensively used to this purpose. These techniques can reduce the dimensionality of large sets of data, increasing their interpretability while minimizing information loss, revealing the underlying correlation structure between the process variables over their time evolution. They do this by projecting the data onto a new set of uncorrelated variables (called principal components) that summarize the original data, in such a way that an intuitive visual comparison of the data evolution patterns across different batches can be obtained.
In this study, we exploit PCA to find the root-cause determining a large variability in the time duration of a key reaction step for an industrial batch process manufacturing a specialty chemical. Large (and unexplained) average batch length and length variability in this reaction step caused significant energy and raw materials consumption per unit of product manufactured.

4.2 Process description

The process under investigation consists in the synthesis of an intermediate chemical for the manufacturing of a hindered amine light stabilizer (HALS), to be used as a polymer additive.
A detailed description of the process is given in Section 2.2, while a piping and instrumentation diagram is shown in Figure 2.4. In this semi-batch process, reactants A and B are fed to the jacketed stirred tank reactor R1. Vacuum is then applied to the system, and the main reaction is a liquid-phase, thermally activated, exothermic one, according to (2.1), where D is the desired product, and C is a byproduct. The process is highly automated and is operated through a recipe that consists of the following steps:

1. setup: the system is set up for a new run;
2. reactant loading: liquid reactants A and B are loaded into R1 from their respective storage tanks in stoichiometric amounts. A small amount of C is also loaded into the reactor; the reason for this is to speed up the initial reaction phase. The reactor is heated up with low-pressure steam through a jacket, to reach the temperature required to carry out the reaction;
3. reaction: vacuum is applied to the system in two steps: a faster pressure decrease is applied first, down to an assigned pressure; then, pressure is further slowly decreased to the pressure value required by the reaction. Once the reaction conditions are met, byproduct C is released as a vapor, and it is removed from the reactor, condensed in C1 condenser, and then stored in T10 buffer tank;
4. nitrogen blanketing: when the required volume in T10 is obtained, vacuum is broken, and the reactor is blanketed with nitrogen;

5. product discharge: liquid byproduct C is discharged from T10 to the waste unit, reaction product D (liquid) is discharged from R1 and fed to the subsequent processing step.

Step 3 corresponds to the reaction phase (the key one for this process), and we call "batch length" its duration. Figure 4.1 shows the distribution of batch lengths recorded over a period of 12 consecutive months before this study was started. It can be seen that the distribution is bimodal, with one mode with a peak at 56 min and one mode with a peak at 74 min; furthermore, the batch lengths range between 55 and 145 min, and the overall average batch length is 77 min.

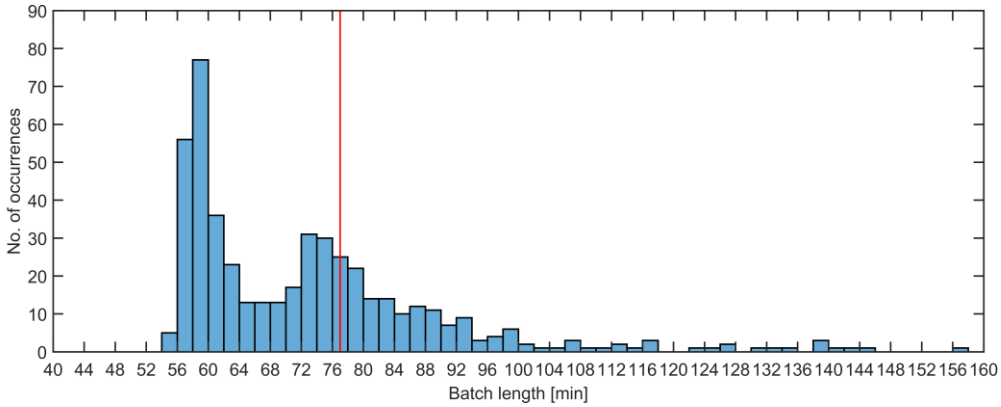


Figure 4.1: *Distribution of the time duration of step 3 in reactor R1 across the historical dataset. The vertical red line is the mean of the distribution.*

The result of this variability in batch length is a decrease in productivity, as well as an increase in energy consumption per unit of product manufactured. Since engineering understanding was not enough to find the root cause of this variability, analytics on the historical manufacturing data was done to mine process-relevant information that could help in the task of troubleshooting the reaction step.

4.3 Available data

A total of $I = 468$ historical batches were extracted from the plant historian (Figure 4.1). All batches ended up in a product meeting the target quality profile. The available data consists of the time trajectories of 7 operating variables, as listed in Table 4.1.

Table 4.1. *Available variables in the* $\underline{\mathbf{X}}$ *dataset*

Variable no.	Variable description
1	R1 absolute internal pressure
2	R1 internal pressure controller output
3	R1 internal temperature
4	C1 condensed liquid temperature
5	T10 internal level
6	R1 internal level
7	Time

The measurements were collected every 30 s.

The available data were organized in a tensor $\underline{\mathbf{X}}[I \times J \times \overline{K}]$, where $J = 7$ is the number of measurement sensors available for the unit, and $\overline{K}$ is the total number of observations per batch, ranging between 183 and 1225.

4.4 Analysis of historical batch data

The dataset was aligned with the indicator variable technique applied to each operating phase, using the internal level as the indicator variable for phases 2 and 5, and time as the indicator variable for phase 3, thus obtaining $\underline{\mathbf{X}}_a[468 \times 7 \times 296]$. Due to their very short time duration, operating stage 1 and 4 were neglected. A BWU MPCA model with 3 PCs was calibrated utilizing the preprocessed dataset $\underline{\mathbf{X}}_a$. The model explains 51% of the dataset variability.

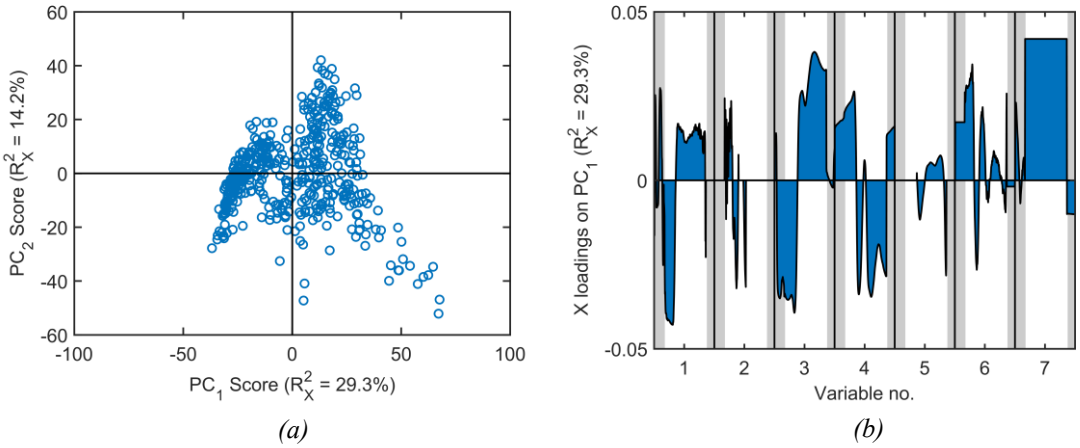


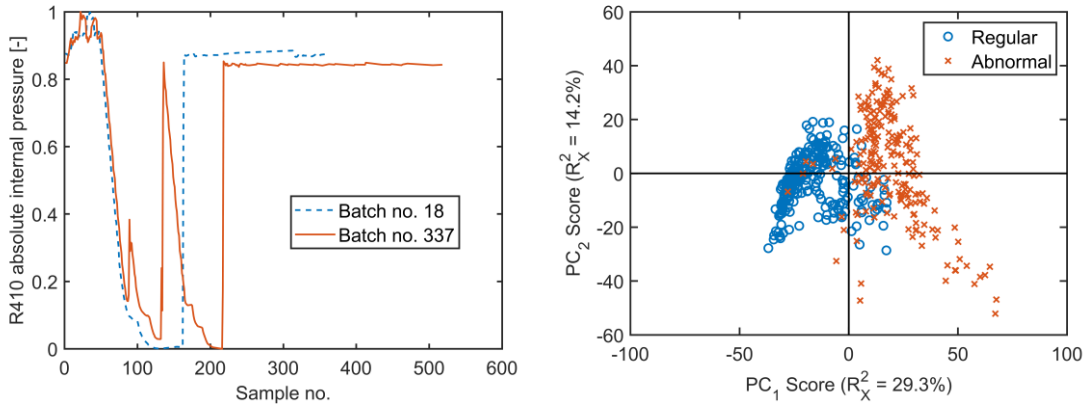
Figure 4.2. *Results of the multiway principal component analysis (MPCA) model built on the historical process data: (a) scores plot, (b) loadings plot on PC1. The grey areas shown for each variable in (b) refer to step 2 (left) and step 5 (right) of the manufacturing recipe; the white area refers to step 3.*

The model scores for the first two PCs are plotted in Figure 4.2a: no particular pattern across the historical batches is apparent. On the other hand, analysis of the loadings (Figure 3b) reveals interesting information. The plot shows how the loadings along the first PC evolve with the aligned time for each of the variables listed in Table 4.1. For any given variable, the time evolution is marked using three background colors: *i*) a grey left area, corresponding to the reactant loading phase (step 2); *ii*) a white central area, corresponding to the reaction phase (step 3); and *iii*) a grey right area, corresponding to the product discharge phase (step 5). Paired analysis of the loadings and scores plot (Bro and Smilde, 2014) helps understanding how each variable concurs to separating the scores along a particular direction in the scores plot. For example, considering the direction along the first PC in the scores plane (left to right in Figure

4.2a), we can conclude from Figure 4.2b that the batches located in the right-half plane are characterized by:

- lower pressure in the first part of the reaction phase, and higher pressure in the second part of the reaction stage (variable no. 1);
- lower temperature in the first part of the reaction stage, and higher temperature in the second part of the reaction stage (variable no. 3);
- lower temperature of the condensed liquid in the condenser in the reaction stage (variable no. 4);
- longer duration of the reaction stage (variable no. 7).

Figure 4.3a reports the time profiles of the reactor internal pressure for two historical batches, respectively projecting onto the left-half plane (batch no. 18) and the right-half plane (batch no. 337) of the scores plot.



(a) *(b)*

Figure 4.3: *(a) Internal pressure profiles from a batch projecting onto the left-half plane of Fig. 4.3a (batch no.18) and a batch projecting onto the right-half plane of Fig. 4.3a (batch no.337). (b) MPCA model scores of Figure 4.2a designated according to whether they are regular or abnormal.*

It is observed that batch no.18 internal pressure follows the recipe described in Section 4.2 correctly, while batch no.337 shows an abnormal pressure profile, with two spikes in the first quarter of the batch duration. This indicates that vacuum was broken (and then reinstated) in that batch before reaching the end of the reaction stage. Confirming the conclusions drawn from the loadings analysis, batch no.337 have a longer duration than batch no.18 (Figure 4.3a), and higher internal pressure during the reaction phase (namely, in-between the initial decreasing ramp and the final step increase). Since loss of vacuum is an abnormal event, an algorithm to automatically identify all historical batches with a reactor pressure profile qualitatively similar to the one of batch no.337 was developed. The relevant batches were denoted as “abnormal”, to distinguish them from the “regular” ones, where the vacuum breakage event did not occur. It was found that the abnormal batches amount to as much as 40% of the historical batches.

The regular and abnormal batches were then identified in the scores plot, obtaining the results of Figure 4.3b. We notice that most of the abnormal batches are projected onto the right-half plane, whereas most of the regular ones are lying in the left-half plane, i.e., the separation between regular and abnormal batches occurs along the first PC. Therefore, we conclude that the difference between regular and abnormal batches acts as the strongest source of variability within the historical dataset.

Discussion with the process experts revealed that the abnormal pressure profiles observed in the historical dataset are related to the intervention of the reactor safety interlock system. In fact, the occurrence of particular combinations of operating conditions in the reactor can trigger the intervention of specific interlocks, each of which acting by breaking the vacuum and blanketing the reactor with nitrogen. The triggering of all potential interlocks was therefore monitored through a dummy variable across a set of new batches, and this enabled the identification of one interlock that did not work properly. This specific interlock was therefore reconfigured, and a new campaign of batches was initiated to validate the finding and assess its impact on the distribution of the batch length across the campaign.

4.4.1 Validation

After reconfiguration of the reactor safety interlock system, 8 months of operating data (corresponding to 635 new batches) were collected from the data historian. Less than 2% of the batches of the validation campaign resulted to be affected by a vacuum loss event, thus confirming that the interlock system reconfiguration was effective. In terms of distribution of the batch lengths across the validation campaign, Figure 4.4 shows that a unimodal distribution is obtained, with a peak value of 58 min. The distribution is also narrower, ranging between 49 and 120 min. Therefore, reconfiguration of the safety interlock system resulted in a 29% reduction in average batch length.

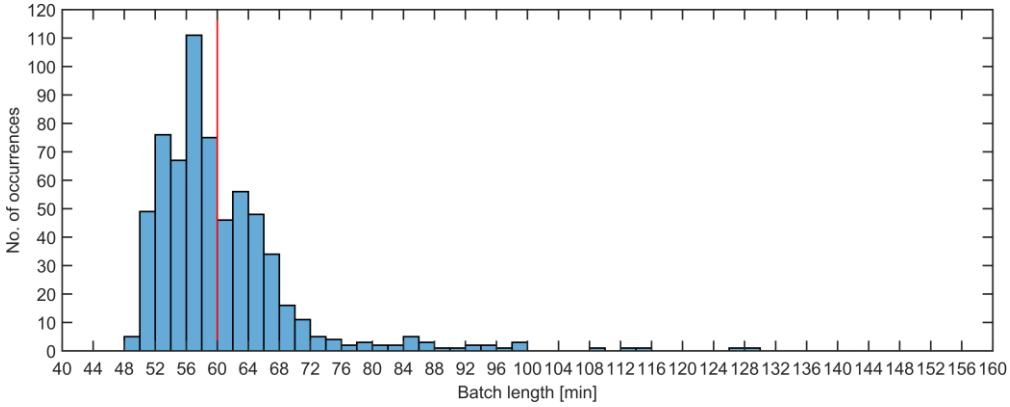


Figure 4.4. *Distribution of the time duration of step 3 in reactor R1 after reconfiguration of the safety interlock system. The vertical red line is the mean of the distribution.*

Overall, this amounted to an 8% reduction of the overall cycle time, with a related saving in the energy expenses. Furthermore, it was estimated that an abnormal batch requires 25% more nitrogen than a regular batch.
Considering that the abnormal batches were 40% of the total number of the historical batches, and assuming that only 2% of the batches of new production campaigns remain abnormal, reconfiguration of the safety interlock system also resulted in a 11% saving on nitrogen expenses.

4.5 Conclusions

Analytics on historical data for this semi-batch manufacturing process, coupled to engineering understanding on the manufacturing process, allowed us to uncover the existence of an abnormal behavior in 40% of the batches in historical manufacturing campaigns. The abnormal batches did not terminate unsuccessfully, but simply took longer to complete than the other batches. Since this did not have an impact on the product quality, they went almost unnoticed, and therefore did not trigger any specific action. However, the longer duration increased the energy expenditure (hence, the operating expenses) per unit of product manufactured. Data analytics was central to find that the cause of the abnormal behavior was the anomalous intervention of one interlock in the reactor safety system, which (under particular conditions) caused vacuum in the reactor to be broken by nitrogen blanketing; this required subsequent vacuum reinstatement to recover the process conditions and to end the batch successfully. Reconfiguration of the safety interlock system allowed us to shorten the average batch length by 29%, and the overall process cycle time by 8%. Furthermore, an 11% reduction on the nitrogen consumption was obtained.

Chapter 5

Development of automated approaches for batch alignment and phase partition for batch end-point estimation*

In this chapter a methodology for automating two challenging steps in the preprocessing of batch data in the development of batch-end point estimation models (soft sensors or classifiers), namely batch alignment and phase partition is presented. The proposed methodology aims at aiding the practitioner at carrying out batch alignment and phase partition with the aim of maximizing the performance of the developed models. Two processes are considered as test beds for the proposed methodology, a real industrial process for the production of an intermediate for polymer additives and a process for the manufacturing of penicillin simulated through the benchmark simulator Pensim. The performance of the proposed methodology are compared with other state-of-the-art batch alignment methodologies in terms of model performance and computational burden. Furthermore, the sensitivity of the proposed methodology to normal variability in the data and to the number of calibration batches is assessed.

5.1 Introduction

Many high value-added products (e.g. specialty chemicals, (bio)pharmaceuticals, food, semiconductors) are obtained by batch processing. Batch processes are run through a recipe, i.e., a sequence of elementary finite-duration processing steps (such as charge, heat up, stir, react, cool down, hold, discharge). Each step is characterized by a given set of operating conditions, and is typically triggered by the occurrence of events (e.g., enough reactant has been fed; temperature reaches a given value; torque exceeds a threshold). Flexibility is a key characteristic of batch manufacturing: by adjusting (either directly or indirectly) the length of the processing steps, a batch process can accommodate variability in the raw materials, operating conditions, and status of the equipment and of the utilities, thus delivering a product that can meet the assigned quality target. As a consequence, a set of batches is often

* Sartori F., Facco, P., Zuecco F., Bezzo F., Barolo M. (2023), Optimal indicator-variable approach for trajectory synchronization in uneven-length multiphase batch processes. *Ind. Eng. Chem. Res.* **62**, 18511-18525.

characterized by an uneven duration between batches, even if all batches manufactured a product that meets the specification.

From the quality control point of view, most batch processes are run at open loop, meaning that quality is assessed only on the end-product at the end of a batch. Depending on the industrial domain, if an off-spec product is detected the batch may be rejected, or reworked, or progressed with a warning to the downstream process that further processes that product. When product quality cannot be measured conveniently (e.g., because a field sensor is not available, or lab analysis takes long to complete), end-point quality assessment is aided by models, which use time-resolved measurements from the plant sensors to either estimate the end-point product quality (models as soft sensors) or simply discriminate between on-spec and off-spec products (models as classifiers). Multivariate statistical methods, such as projection onto latent stuctures (PLS; Geladi and Kowalski, 1986; MacGregor *et al.*, 1994; Wold *et al.*, 1984), PLS discriminant analysis (PLS-DA; Barker and Rayens, 2003) and their multiway extensions (Nomikos and MacGregor, 1995b), offer convenient modeling environments in this context, because they are interpretable and preserve time resolution in the available data (Rendall *et al.*, 2019).

A crucial aspect when using multiway-PLS or multiway-PLS-DA as modeling platforms for quality assessment is that they typically require data alignment, namely equalization (all the variables are expressed at the same sampling rate across batches) and synchronization (all the landmarks for the variables trajectories are aligned in time across batches (González-Martínez *et al.*, 2018). Ad-hoc synchronization techniques, such as truncating the trajectories of all batches to the shortest batch length (Rothwell *et al.*, 1998) or extending the length of shorter batches by repeating the last measurement (Lakshminarayanan *et al.*, 1996), are simple workarounds that can be set up quickly for preliminary dataset screening and analysis, but may provide ineffective data modeling (Rendall *et al.*, 2019). A more effective, yet still simple, synchronization strategy consists in nonlinearly mapping time to an indicator variable (IV; Nomikos and MacGregor, 1994), namely to a measured variable that *i*) progresses monotonically in time, *ii*) has a favorable signal-to-noise ratio (García-Muñoz *et al.*, 2003; Ündey *et al.*, 2003), and *iii*) has the same starting and ending values for all batches. The IV is to be selected by the user based on process knowledge (García-Muñoz *et al.*, 2011; Kourti, 2003). It may not exist for an entire batch, but can exist for single time windows wherein the measured variables have similar correlation structure. Each such window is called a batch phase (not to be confused with a batch processing step). The IV approach for trajectory synchronization is very popular and proved effective in a number of applications (Barton *et al.*, 2021; Brunner *et al.*, 2020; García-Muñoz *et al.*, 2003; Kourti *et al.*, 1996; Krause *et al.*, 2015; Neogi and Schlags, 1998). However, when several potential IVs exist, it may not be obvious which one is the most appropriate to choose. Furthermore, partitioning a batch into phases is a challenge in itself(Guo and Jin, 2019; Lu *et al.*, 2004; Luo *et al.*, 2016; Zhang *et al.*, 2018),

because phases do not necessarily match the occurrence of physical events in a process (i.e., phases do not necessarily match processing steps). Finally, phase partitioning and batch synchronization have been mostly regarded as two independent activities, despite the fact that they are both functional to the model that needs to be developed. Indeed, both phase partitioning and batch synchronization are known to have a strong impact on the model performance (González-Martínez *et al.*, 2018; Zhao, 2014).

Advanced synchronization techniques exist that do not use an IV for synchronizing batch trajectories. Dynamic time warping (DTW; Kassidas *et al.*, 1998), correlation optimized warping (COW; Fransson and Folestad, 2006; Nielsen *et al.*, 1998), and multisynchro (MS; González-Martínez *et al.*, 2014) are the most popular among them. DTW synchronizes two trajectories by translating, compressing and expanding them so that similar features within them are matched. The method is inherently multivariate, since it does not rely on a single variable to perform the synchronization. However, it requires selecting a reference batch the synchronized ones should be matched to; furthermore, its computational burden scales badly with the dataset size (namely, with the number of time points characterizing each trajectory; Zhou and Wong, 2008). Finally, DTW is known to generate artifacts when some batches are significantly shorter than the chosen reference (José M. González-Martínez *et al.*, 2014). COW is based on maximizing the correlation between two trajectories, and is less computationally demanding than DTW. However, it is univariate by design, because each variable is synchronized separately from the others. Moreover, it can generate artifacts, and requires identifying a reference batch and using it for synchronizing all other batches (Lu *et al.*, 2016). MS aims not only at minimizing a defined distance between a reference batch and the other batches in a dataset, but also at removing particular asynchronous behaviors that it can identify among batches (e.g., incomplete batch runs; delayed measurement collection; natural variability). Arguably, it is the most advanced batch synchronization algorithm proposed to date, it is based upon DTW, and can therefore be computationally intensive.

It is to be noted that some multivariate statistical techniques, such as PARAFAC2 (Luo *et al.*, 2016) and GHOPLS-CP (Luo *et al.*, 2015), can deal with time-resolved data from uneven-length batch processes without the need for batch synchronization. However, these techniques are computationally inefficient (Tian *et al.*, 2018; Yu *et al.*, 2021; Zhang *et al.*, 2018) and more sensitive to noise (Amigo *et al.*, 2008) with respect to multiway PLS and multiway PLS-DA. On the other hand, when retaining time resolution is not a requirement, one can resort to feature-oriented data analysis (He and Wang, 2011; Rato *et al.*, 2017; Rendall *et al.*, 2017a), which does not require batch synchronization.

In this study, we propose an optimal IV approach (IVopt) for trajectory synchronization in uneven-length multiphase batch processes. The ultimate aim is building an effective multivariate statistical model for end-point quality assessment once a batch is terminated. The idea behind IVopt is that phase partition and trajectory synchronization are carried out

simultaneously, rather than disjointly, using an optimization framework based on surrogate modeling, with the aim of maximizing the performance of the product quality assessment model that is under development. Within this approach, we resort to surrogate optimization in order to find the optimal phase partition parameters, and we propose a novel methodology for automatic identification of the most appropriate IV within each batch phase. We challenge IVopt against standard and advanced synchronization strategies, namely trajectory truncation (TR), trajectory extension (EXT) with mean values, IV (with *a-priori* phase partitioning and IV selection based on engineering judgment), DTW, COW, and MS. We use two case studies as test beds: an industrial fed-batch process for the manufacturing of a specialty chemical, and a simulated fed-batch process for the manufacturing of penicillin (Birol *et al.*, 2002).

5.2 Proposed optimal indicator-variable synchronization methodology

Figure 5.1 shows a flow-chart of the proposed IVopt approach for trajectory synchronization in uneven-length multiphase batch processes. The methodology iteratively adjusts the set $\boldsymbol{\xi} = [N, \Lambda, V]$ of parameters defining the partitioning of the available batches into phases until the resulting quality assessment model is optimal in some sense (to be discussed later). The distinguished features of the IVopt algorithm are the following:

- phase partition is targeted to optimal model-based quality assessment; namely, phase partition, batch synchronization and quality assessment are not disjoint activities;
- the most appropriate IV for trajectory synchronization within each phase is identified automatically;
- a surrogate optimization approach is used to iteratively update the phase partition parameters.

Eventually, IVOpt returns both the optimal set $\boldsymbol{\xi}_{\text{opt}}$ of phase partition parameters, and the most appropriate IVs to be used for batch synchronization. New batches can then be synchronized using this information. Next, we discuss the main steps along which the proposed methodology develops.

5.2.1 Automatic phase partition revisited

Automatic phase partition is done following the methodology illustrated in Section 3.1.5.2. However, we found that the use of a switch control limit as defined in (3.26) may suffer from noise when the signal-to-noise ratio is not high enough. To attenuate the impact of noise in phase identification, the information from the values of δ_k calculated after Λ movements of the moving window is included in the calculation of an adaptive control limit:

$$\sigma_k = \frac{1}{k}\sum_{z=1}^{k} \delta_z \quad , \tag{5.1}$$

which therefore includes not only the last Λ calculated values of δ_k, but all the values from the beginning of the current phase. Upon application of the phase partition methodology to all batches in $\underline{\mathbf{X}}_{\mathrm{B}}$, the optimal set $\boldsymbol{\xi}_{\mathrm{opt}}$ of phase partition parameters is found, from which the distribution of the number of phases identified across all batches is obtained. The mode of this distribution is set as the actual number $\bar{\Phi}$ of phases to be used for all batches. If, for a given batch i, the number of identified phases is $\Phi_i \neq \bar{\Phi}$, then that batch is forced to partitioning into $\bar{\Phi}$ phases by assigning phase switch time points equal to the average of the phase switch time

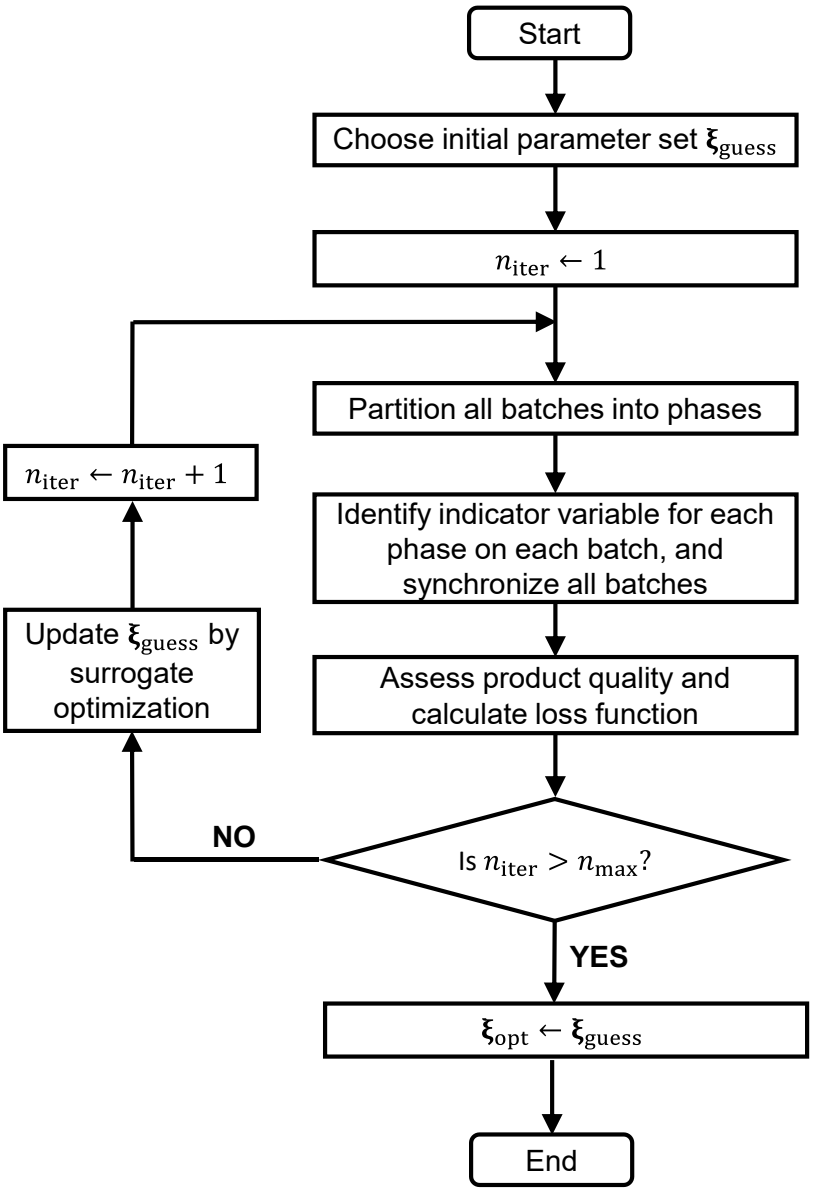


Figure 5.1. *IVopt: flow chart of the proposed optimal indicator-variable approach for trajectory synchronization in uneven-length multiphase batch processes.*

points obtained for all the batches for which $\Phi_i = \bar{\Phi}$ is found. Note that this typically occurs for a limited number of batches only.

5.2.2 Automatic indicator variable identification and batch synchronization

We propose a methodology that automatically returns an appropriate IV for each of the $\bar{\Phi}$ phases identified across the entire $\underline{\mathbf{X}}_B$ dataset. For a given phase, the methodology identifies a given process variable as a candidate IV if it simultaneously fulfills the following three conditions: *i*) it is monotonic, *ii*) it has a sufficiently high signal-to-noise ratio, and *iii*) it has approximately the same initial and final values across all batches in $\underline{\mathbf{X}}_B$. Next, we discuss how fulfillment of the conditions is assessed for a given phase and a given variable.

Condition *i*) is assessed by performing a Mann-Kendall test for monotonicity (Gilbert, 1987). The test returns a Yes/No condition (at 0.05 significance level) to the null hypothesis that the trend of the given variable is non-monotonic.

In order to assess if condition *ii*) is fulfilled, the variable is first detrended by subtracting the best straight-line fit from the variable (as implemented in Matlab R2020a); then, the standard deviation of the detrended variable is compared to the range of the non-detrended one: if the standard deviation is smaller than the range, then the signal-to-noise ratio of the variable is deemed acceptable. An F-test (at 0.05 significance level) is carried out to verify that the variance of the detrended variable and the variance of the non-detrended variable are statistically different.

Finally, condition *iii*) is met for the variable under investigation if both the following inequalities are satisfied: $R > \sigma_i$ and $R > \sigma_e$, where R is the range of the variable values in the phase, and σ_i and σ_e are the standard deviations of the phase initial and end points, respectively.

The actual IV among all the identified candidate IVs for a given phase is selected as the one for which the Mann-Kendall test is satisfied more strongly (smallest average p-value across all batches). If no process variable is identified as a candidate IV using the above approach for a given phase, time is used as the IV for that phase, since time always fulfills the first two conditions, and in most cases also the third one (at least to some approximation).

Once an IV is identified for each phase, the batches are synchronized in a phase-by-phase fashion using the IV approach (García-Muñoz *et al.*, 2003).

5.2.3 Product quality assessment and loss function calculation

Product quality assessment is done through a soft sensor or a product classifier by building a PLS or a PLS-DA model (respectively) on the synchronized batches. The relevant loss functions in 10-fold cross validation ($RMSECV$ and $(1 - Accuracy)$, respectively) are then calculated.

5.2.4 Phase partition parameters update by surrogate optimization

As noted earlier, the resulting phase partition (hence, the performance of the quality assessment model) strongly depends on the set $\boldsymbol{\xi} = [N, \Lambda, V]$ of parameters used within the phase partition methodology. The IVopt algorithm optimizes the selection of $\boldsymbol{\xi}$ by minimizing the loss function $\mathcal{L}(\boldsymbol{\xi})$ associated to the quality assessment model. The optimization problem can be formulated as

$$\min_{\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{\xi})$$

$$\text{subject to:} \tag{5.2}$$

$$lb_N \leq N \leq ub_N \ , \quad lb_\Lambda \leq \Lambda \leq ub_\Lambda \ , \quad lb_V \leq V \leq ub_V$$

where lb_z and ub_z denote the lower bound and the upper bound of parameter z.
We solve the optimization problem using surrogate optimization (Gutmann, 2001). Further details on surrogate optimization can be found in Section 3.2.

5.3 Case studies

Two case studies are considered to test the proposed IVopt framework: an industrial batch process for the manufacturing of a specialty chemical, and a simulated fed-batch process for the manufacturing of penicillin. Next, we provide details about them.

5.3.1 Case study #1: industrial fed-batch manufacturing of a specialty chemical

Figure 2.5 shows the piping and instrumentation diagram of the industrial fed-batch process under investigation (Reaction 2 in Figure 2.2), where product *P* (actually, an intermediate used in the manufacturing of a polymer stabilizer) is obtained in jacketed reactor R5 (6.5 m^3 volume) from the following catalytic reaction (2.2), where species B is a liquid reactant, species D is a gaseous reactant, and G is the desired species. Product *P* is mainly made of G, traces of unreacted B and other subproducts. The manufacturing recipe is quite complex, and can be summarized by the following finite-length operating steps:

- Reactor R5 is set up for a new batch.
- Reactant B and catalyst are loaded into R5.
- R5 is blanketed with nitrogen.
- Reactant D is fed to R5 and pressurizes it until an assigned pressure is reached; after that, the feed is stopped, and the reaction is allowed to proceed for an assigned amount of time.

The profile through which B is fed depends on several factors and is quite complex, resulting in a very strong variability of this phase.

- R5 is vented.
- R5 is blanketed with nitrogen.
- Product P is discharged from R5 to a downstream plant section, where it is further processed.

Too large an amount of unreacted B in *P* can be an issue for quality, because *P* is used as a reactant in a downstream unit, and an excess of B can downgrade the optical properties of the final product. A lab assay of *P* is taken for some batches only. Real time measurements of some process variables are available as listed in Table 5.1.

Table 5.1. *Case study #1: variables measured in real time.*

Variable no.	Variable name
1	Totalized reactant D fed
2	Reactant D flow rate
3	Reactant D flow rate controller output
4	R1 internal absolute pressure
5	R1 internal pressure controller output
6	R1 internal pressure controller 2 output
7	R1 internal temperature
8	R1 internal temperature controller output
9	R1 internal temperature difference controller output
10	Time

From the data historian, a set of 52 batches (completed across years 2020 and 2021) are collected for which the end-point quality (in terms of concentration of B in *P*) is measured. Within this dataset, the batch length ranges between 7 h and 19 h. The measured variables are downsampled to one every 2 min. The dataset is split in 36 calibration batches (24 batches ending up in a “good” product, and 12 batches ending up in a “bad” product) and 16 validation batches (10/6 good/bad).

The quality assessment model is required to classify the quality of product *P* as either good or bad, once a batch is come to an end, using the time-resolved measurements of the variables listed in Table 2. A multiway PLS-DA model is developed to this purpose.

5.3.2 Case study #2: simulated fed-batch manufacturing of penicillin

We consider a fed-batch fermentation process that manufactures penicillin. The process is simulated using Pensim (Birol *et al.*, 2002), a software used in several process control and monitoring studies (Birol *et al.*, 2002; Reis *et al.*, 2021; Wan *et al.*, 2014).

Figure 5.2 shows a simplified piping and instrumentation diagram of the process.

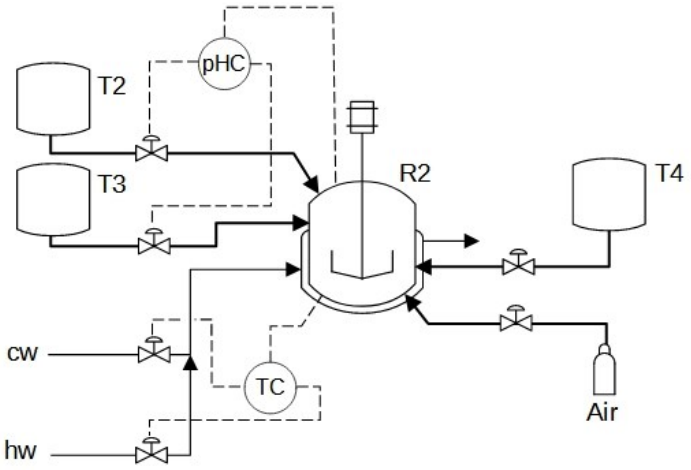


Figure 5.2 *Case study #2: simplified piping and instrumentation of the simulated process for the manufacturing of penicillin*

The penicillin manufacturing recipe is based on two processing steps:

1. A batch culture step, where the reactor is initially loaded with Penicillium Chrysogenum and glucose from tank T4, and the reaction starts; this step ends when the concentration of glucose in reactor R2 drops below an assigned threshold;
2. A fed-batch step, where pH is automatically controlled through the addition of acid from tank T2 and base from tank T3; during this step, glucose and air are fed constant rates. The end-point condition is reached when the total volume of glucose fed to R2 during this step reaches 14 L(Sun *et al.*, 2011).

Real time measurements of some process variables are available as listed in Table 5.2.

Table 5.2 *Case study #2: variables measured in real time.* $stdev$ *is the standard deviation of a zero-mean normal distribution of random numbers.*

Variable no.	Variable name	Units	$stdev$
1	Dissolved oxygen	g/L	0.0067
2	Bulk volume	L	0.033
3	pH	[-]	0.0167
4	Temperature	K	0.17
5	Glucose feed rate	L/h	0.17
6	Aeration rate	L/h	0.0834
7	Agitator power	W	0.17
8	Glucose feed temperature	K	0.17
9	Jacket water flow rate	L/h	0.83
10	Cumulated base flow	L	$3.33 \cdot 10^{-6}$
11	Cumulated acid flow	L	$3.33 \cdot 10^{-7}$

Measurement noise is simulated in the form of additive random numbers sampled from a normal distribution with zero mean and standard deviation $stdev$ as indicated in Table 5.2 *(*Vanlaer *et al.*, 2012). Process variability is generated by randomly changing the values of some initial conditions and some operating variables, as detailed in Table 5.3. Further variability is

generated by assuming that the threshold glucose concentration determining the switch between operating steps 1 and 2 randomly varies between 0.3 and 7 g/L.

Table 5.3 *Case study #2: nominal initial conditions, nominal operating variables, and variability around them (ϵ is sampled from a standard normal distribution).*

Initial condition	Units	Nominal value
Glucose concentration	g/L	$15 + \epsilon$
Dissolved oxygen	%	1.16
Biomass concentration	g/L	0.1
Penicillin concentration	g/L	0
Culture volume	L	$150 + 10\epsilon$
CO_2 concentration	mmol/L	$0.75 + 0.05\epsilon$
Hydrogen ion concentration	mol/L	$10^{-5+0.1\epsilon}$
Fermentor temperature	K	298
Generated heat	kcal/h	0
Operating variable	**Units**	**Nominal value**
Aeration rate	L/h	8
Agitator power	W	$30 + \epsilon$
Glucose feed rate	L/h	$0.04 + 0.0025\epsilon$
Glucose feed temperature	K	296
Culture volume	L	$150 + 10\epsilon$
pH	[-]	5
Fermentor temperature	K	298

A set of 300 batches is generated. Within this dataset, the batch length ranges between 345 and 479 h. The variables measured in real time are sampled every 0.5 h. The dataset is split in 250 calibration batches and 50 validation batches. The quality assessment model for this case study is required to estimate, at the end of a batch, the end-point penicillin concentration using the time-resolved measurements in Table 5.2. A multiway PLS model is developed to this purpose.

5.4 Results

We benchmark IVopt against DTW, COW, TR, EXT, IV, and MS for the two case studies illustrated in the previous section. The software used for carrying out DTW and MS is the MVBatch toolbox (González-Martínez *et al.*, 2018), the software implementing the other methodologies has been developed in-house. When applying the IV, IVopt, DTW, MS and COW methodologies, the warping profile is appended to the synchronized dataset as it contains relevant information (García-Muñoz *et al.*, 2003). The batch synchronization methodologies are compared in terms of performance of the relevant quality assessment model and computational cost. For a given batch, IVopt is applied using the automatic phase partition methodology as discussed in Section 5.2.1; DTW, COW, MS, TR and EXT are applied without any phase partition, and IV is applied by partitioning a batch into processing steps rather than

into phases. The computation time refers to the use of a laptop computer equipped with an Intel Core i7-9750H 2.60GHz CPU and 32 GB of RAM.

5.4.1 Results for case study #1

With reference to IVopt, the initial phase partition parameter guesses are $N = 1.08$, $\Lambda = 11$, and $V = 10$, and the following constraints are enforced in the surrogate optimization algorithm: $1 \leq N \leq 4$, $2 \leq \Lambda \leq 15$, and $5 \leq V \leq 35$. The maximum number of iterations is set to 100. The algorithm returns a 10-fold cross-validation accuracy of 92% (with 2 latent variables) in the calibration dataset at $N = 2.3$, $\Lambda = 5$, and $V = 15$.

Using the optimal phase partition parameter set, the classification accuracy for the validation dataset is 100%, meaning that all validation batches are classified correctly. A comparison of the product quality classification results obtained for the validation dataset for the batch synchronization methods considered in this study is shown in Figure 5.3 (the reported optimal number of latent variables is determined by cross-validation using the calibration dataset). IVopt outperforms all other synchronization methodologies except IV (which, however, requires assigning manually both the phase partitioning and the indicator variable within each phase). Some synchronization methods (namely EXT and COW) lead to poor classification accuracy. It is somewhat surprising that DTW is performing better than MS, given that, according to the MS algorithm, DTW alignment is selected when this is the most appropriate method to apply.

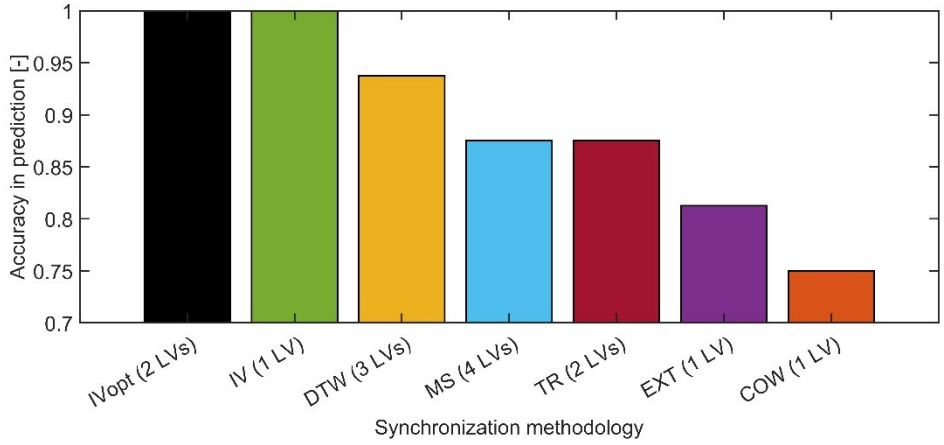
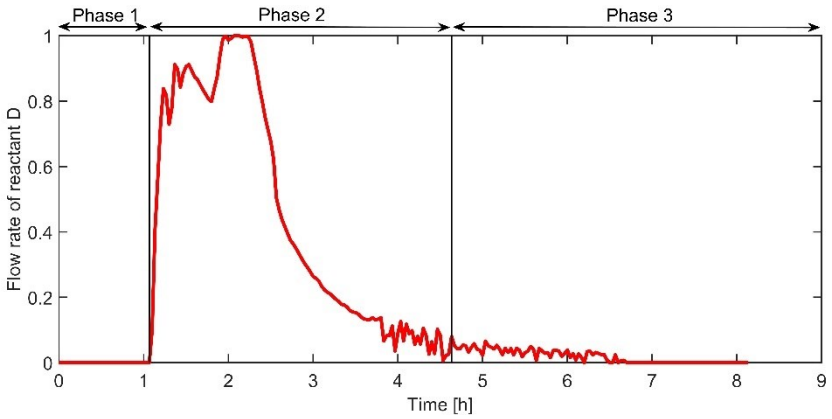


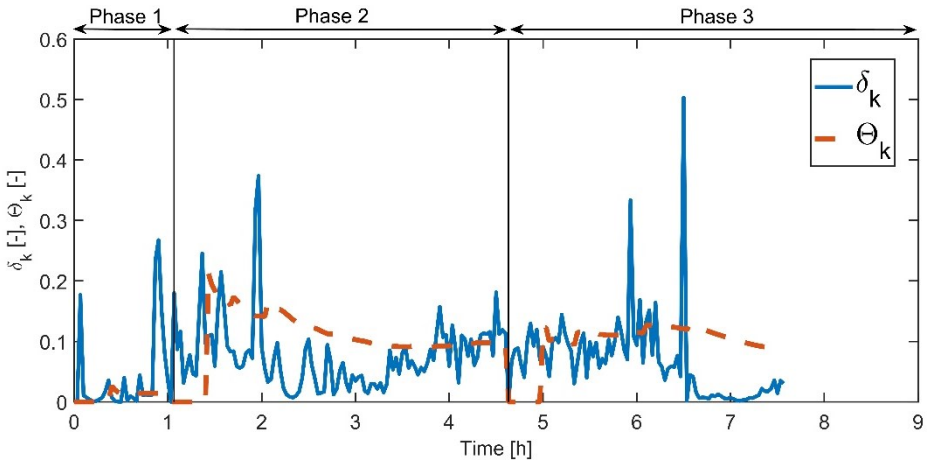
Figure 5.3 *Case study #1. Classification accuracy obtained in prediction using the validation dataset for different batch synchronization methodologies (the numbers in parenthesis indicate the optimal number of latent variables as determined by cross-validation using the calibration dataset).*

Despite the fact that as many as 7 operating stages exist, IVopt returns a batch partitioning into only three phases. Figure 5.4a shows the time profile of the flow of reactant D to reactor R5 for a representative batch together with the phase partitioning returned by IVopt. The partitioning

is physically meaningful: the phase switching points correspond roughly to the time points when the flow rate of D starts to be greater than zero, and then returns close to zero. The automatically selected indicator variables are time for phase 1, and the totalized amount of reactant D fed to R5 for both phase 2 and phase 3. The time profiles of the gain index δ_k and its threshold value Θ_k are illustrated in Figure 5.4b. It can be seen that both of them get adjusted as the batch progresses; when δ_k values are consistently greater than the corresponding values of Θ_k, a phase switch occurs.



(*a*)



(*b*)

Figure 5.4 *Case study #1, representative batch. (a) Phase partition obtained by the IVopt methodology: time profiles of the (dimensionless) flow rate of reactant D; (b) time evolution of the δ_k and Θ_k parameters.*

Figure 5.5 compares the computer time required to carry out batch synchronization for all methods. Although the time required by IVopt is significantly greater than the one required by any of the other methods, it is nevertheless very short (5 min). Computationally demanding methods like DTW and MS require less than 1 min to run (hence, much less than IVopt), because the dataset to be analyzed has quite a small size (on average, ~250 samples per

measured variable in a batch). However, these methods scale badly with the dataset size, as will be shown for case study #2.

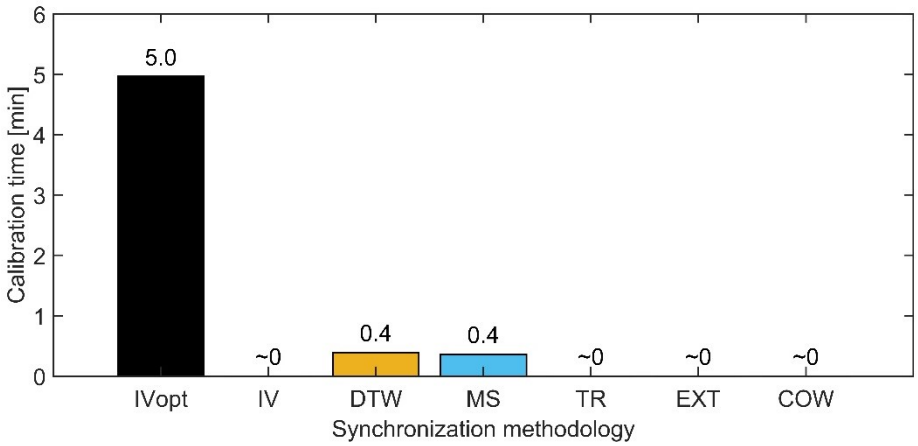


Figure 5.5 *Case study #1. Time required to carry out the synchronization of the calibration dataset trajectories using different techniques.*

The synchronization results are illustrated in Figure 5.6 for all synchronization methods, with reference to the profiles of the flow rate of reactant D (variable no. 2) across the calibration batches. The unsynchronized trajectories are shown in Figure 5.6a. It can be seen that COW (Figure 5.6b) struggles to obtain an effective synchronization for some batches. DTW (Figure 5.6c) and MS (Figure 5.6d) effectively minimize the differences between the trajectories. However, to achieve this they introduce distortions in some trajectory segments, particularly when a strong compression is applied; these distortions appear as horizontal segments for several trajectories, approximately located between synchronized time 50 and 100. IV (Figure 5.6e) and IVopt (Figure 5.6f) work differently from the other synchronization methods. Recall that IVopt selects the totalized volume of reactant D as the indicator variable during phase 2 and phase 3. This indicator variable is basically the time integral of the variable shown in Figure 5.6a. Therefore, the portions of a trajectory with larger values of the reactant D flow rate within phases 2 and 3 are expanded in time (thus magnifying the trajectory differences across batches) to maximize the classification accuracy; on the other hand, the portions with smaller values are contracted (minimizing such differences), as they have less impact on the classification. Therefore, when using IVopt (Figure 5.6f), time is substituted with an indicator variable that is nonlinearly related to time itself and is more descriptive of the progress of the process. The IV method (Figure 5.6e) works somewhat similarly to IVopt, resulting in a similar classification performance. Yet, IVopt performs all operations (phase partition and indicator variable selection) automatically.

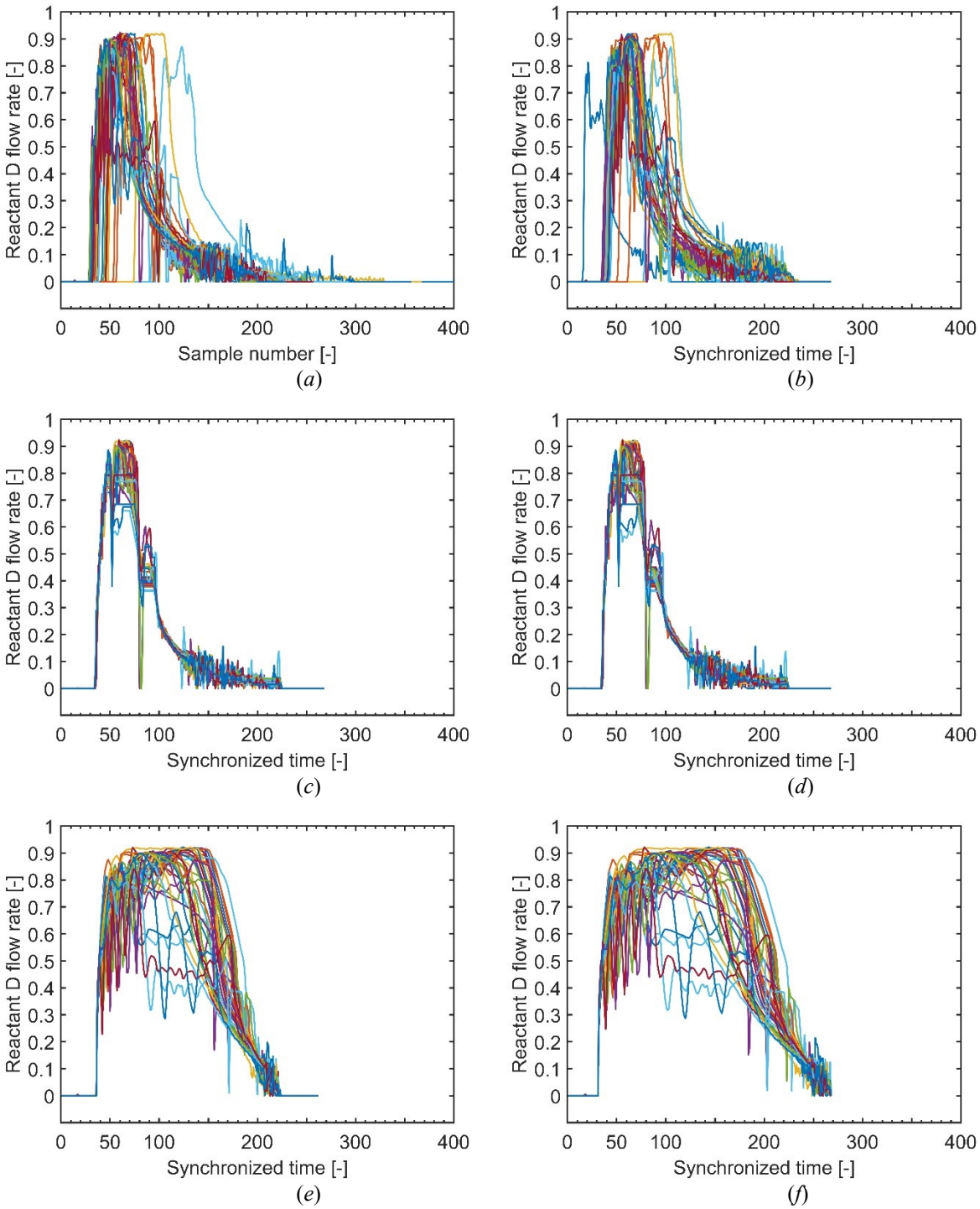


Figure 5.6 *Case study #1. Time trajectories of the (dimensionless) flow rate of reactant D (a) without synchronization, and synchronized using (b) COW, (c) DTW, (d) MS, (e) IV, and (f) IVopt.*

5.4.2 Results for case study #2

The optimization is carried out using $N = 1.8, \Lambda = 2$, and $V = 30$ as initial guesses, and the following box constraints: $1 \leq N \leq 3$, $1 \leq \lambda \leq 5$, and $5 \leq V \leq 100$. The surrogate optimization algorithm iterates 300 times yielding the following optimal parameter set: $N = 1.02, \Lambda = 1$, and $V = 98$.

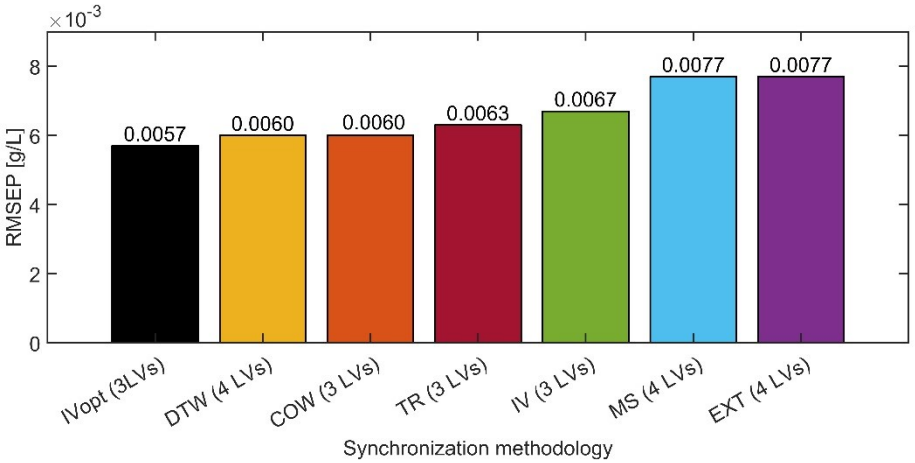


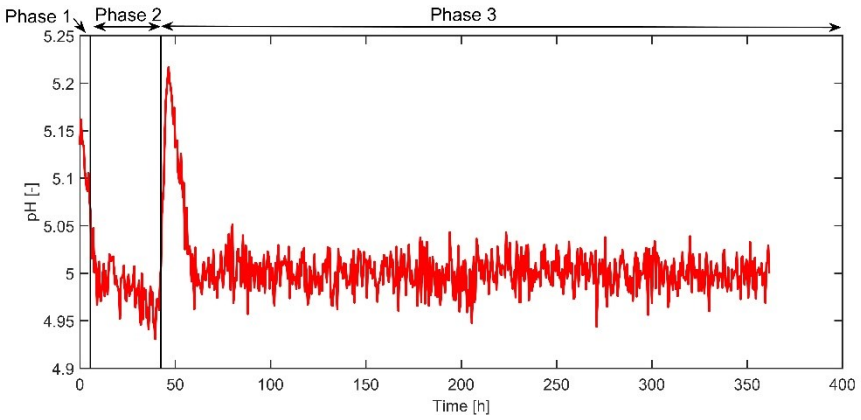
Figure 5.7 *Case study #2. Root-mean squared prediction error using the validation dataset for different batch synchronization methodologies (the numbers in parenthesis indicate the optimal number of latent variables as determined by cross-validation using the calibration dataset).*

Figure 5.7 shows that IVopt provides the best validation results among all synchronization methods: the root-mean squared error of prediction is 0.0057 g/L, slightly better than with DTW and COW, and considerably better than with MS and EXT.

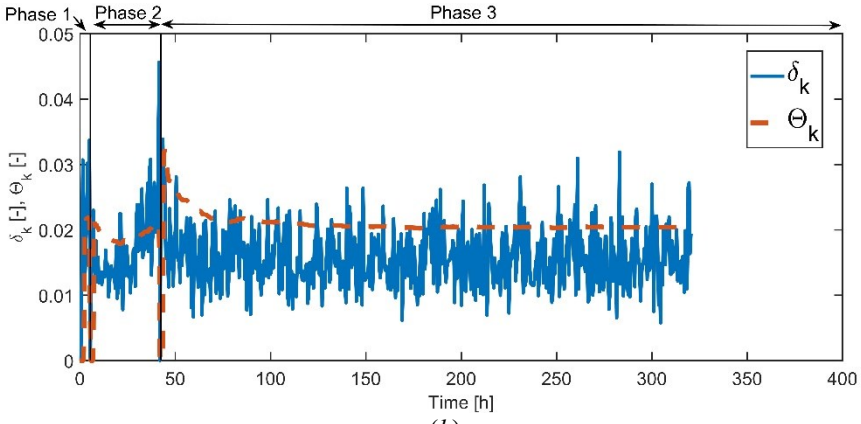
IVopt identifies 3 phases (Figure 5.8a). In the (very short) first phase, time is selected as the indicator variable, whereas in the second and third phases, the cumulated base flow rate is selected as the indicator variable. Figure 5.8 suggests that abrupt changes in pH correspond to large variations in the correlation structure of the measured variables (hence, to phase switch), as captured by the time profiles of δ_k and Θ_k.

Figure 5.9 clarifies that, whereas the computational time required by IVopt is not negligible (~30 min), it is slightly smaller than the one required by DTW (~40 min), and much smaller than required by MS (~39 h).

Comparing Figure 5.5 and Figure 5.9, in the face of a dataset size increase from ~90k data entries (case study #1) to ~2M data entries (case study #2), the computational time required by IVopt increased by ~6 times, while DTW increased by ~100 times and MS by ~5900 times.



(*a*)



(*b*)

Figure 5.8. *Case study #2, representative batch. (a) Phase partition obtained by the IVopt methodology: time profiles of pH; ((b) time evolution of the δ_k and Θ_k parameters.*

Therefore, IVopt scales with the dataset size much better than these other two advanced synchronization methodologies.

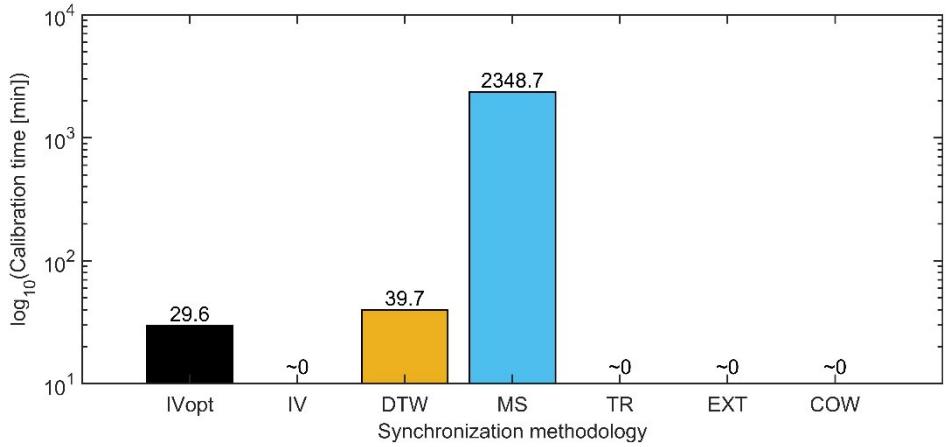


Figure 5.9 *Case study #2. Computation time required to carry out the synchronization of the calibration dataset trajectories using different synchronization techniques.*

We assessed the sensitivity of the model performance to the number of calibration batches for all synchronization methods. Figure 5.10 clarifies that using a smaller number of calibration batches leads to a minor loss of performance for all methods, unless very few (namely, 20) calibration batches are used. However, IVopt outperforms the other methods also in this limiting case, as its root mean square error in cross validation (RMSECV) is the smaller among the tested methodology for each number of calibration batches.

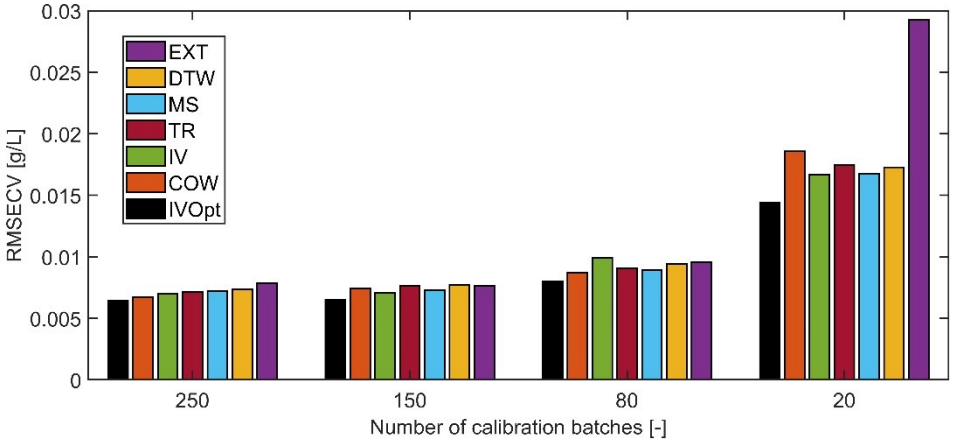


Figure 5.10 *Case study #2. Impact of the number of calibration batches on the root-mean squared prediction error using the validation dataset for different batch synchronization methodologies.*

To evaluate the robustness of the phase partition parameters against the inherent variability in the data, we tested the optimization results on 50 different (random) splits of the available data into calibration/validation datasets. The distribution of the optimal parameters is illustrated in Figure 5.11.

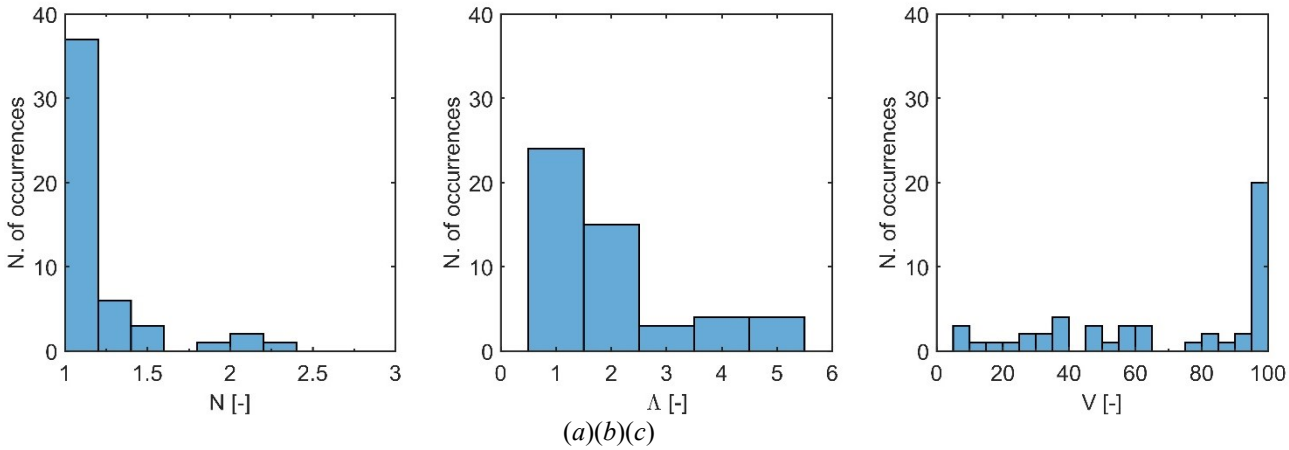


Figure 5.11 *(a) Case study #2. Distribution of the optimal phase portioning parameters for 50 different splits of the calibration/validation datasets: (a) N; (b) Λ; (c) V.*

It is observed that N (a real number) has a very narrow distribution around 1.0. Parameters Λ and V (natural numbers) exhibit slightly greater variability, but in most cases Λ is either 1 or 2,

and V is around 100. We conclude that the results obtained by IVopt are robust to typical fluctuations in the data.

5.5 Conclusions

This paper presented a novel methodology (called IVopt) for phase partitioning and trajectory synchronization in uneven-length multiphase batch processes. The methodology retains the effectiveness of a simple trajectory synchronization methodology like the classic indicator variable (IV) approach, but improves it in two directions; namely, partitioning into phases and selection of the most appropriate IV within each phase *i*) are performed automatically rather than manually, and *ii*) are carried out simultaneously rather than disjointly, based on an optimization framework that maximizes the performance of a model for product quality assessment that is to be built using the available datasets. Differently from classic IV and from advanced synchronization methodologies like dynamic time warping (DTW), correlation optimized warping (COW), and multisynchro (MS), the proposed methodology is process-agnostic, i.e., it does not require identifying either a reference batch or a reference variable.

To test the proposed data preprocessing methodology, we considered two datasets, one from an industrial process and one from a simulated process. We compared the performance of the resulting product quality assessment model when the available data were preprocessed with IVopt and with other synchronization strategies, namely trajectory truncation (TR), trajectory extension (EXT) with mean values, classic IV, DTW, COW, and MS. IVopt always led to the best model performance, both when the model was used as a soft sensor to estimate the product end-point quality, and when the model was used as a classifier to discriminate between on-spec and an off spec products. In this latter case, also classic IV led to excellent classification performance, but partitioning into phases and selection of the most appropriate IV within each phase had to be done manually, based on engineering judgment.

From the computational side, IVopt is more demanding than simple (and less effective) strategies like IV, TR and EXT, and also than COW. Compared to other advanced methodologies, whenever DTW and MS become computationally intensive, IVopt outperforms them, the more so as the number of samples per batch increases.

Chapter 6

Development of guidelines for phase partition and batch alignment-free methodologies*

In this chapter a set of guidelines for the implementation of a phase partition and batch alignment-free methodology are presented, complete of a clear technique for fault detection and a novel approach to fault diagnosis. The methodology, implemented according to the provided guidelines, is compared with a traditional batch process monitoring methodology and the two are compared both in terms of fault detection strength and detection speed over five case studies coming from different industrial sectors.

6.1 Introduction

Batch processing is widespread in the industrial sectors where low volumes of high value-added products are produced and flexibility is a paramount requirement e.g., due to relevant oscillations in the product market demand. Some examples of products include (bio)pharmaceuticals, polymers, semiconductors and food. Batch processes are run through the cyclic repetition of a sequence of elementary processing steps (raw materials charge, heat up, stir, react, cool down, hold, discharge) called recipe, with finite time duration. Each processing step is characterized by a given set of operating conditions and it is triggered by the occurrence of a specific event.

Adjustments to the recipe can be used for accommodating some variability entering the process (e.g. raw materials, utilities), varying the time duration of each batch and generating uneven-length batches. Monitoring process operations is necessary in order to provide a safe operating environment and manufacture a high-quality product. Typically, distributed control systems (DCS) are used for process monitoring, providing the operators a large amount of information on the process in a univariate fashion. It is still difficult to detect abnormal process deviations and make proper and timely decisions to eliminate them as operators can only focus on a few

* Sartori F., Facco, P., Bezzo F., Barolo M. (2023), On the application of assumption free modelling for multivariate statistical batch process monitoring. *In preparation*.

process variables at a time (Ji and Sun, 2022). The intrinsically dynamic and non-stationary nature of batch processes renders this task even harder.

Multivariate statistical process monitoring (MSPM) is a data-driven monitoring approach that aims at reducing the number of monitored variables by summarizing the operating conditions in a small set of latent variables describing the physico-chemical state of the process (Wise and Gallagher, 1996). The main tasks carried out in MSPM are two: (*i*) uncovering that an anomaly is process conditions has happened and (*ii*) uncovering the nature of the observed anomaly; (*i*) is called "fault detection" while (*ii*) is called "fault identification and diagnosis" (Joe Qin, 2003). The modelling activity required for implementing MSPM consists in collecting data from several historical batches in normal operating conditions (NOC), calibrating a model with these data and then use the model to compare these batches with a new batch in an online fashion.

The multiway extensions of principal component analysis (PCA, Jollife and Cadima, 2016) are a commonly adopted modelling methodology for batch process monitoring as they are interpretable and preserve time resolution in the available data (Rendall *et al.*, 2019). Fault detection is carried out comparing the Hotelling's T^2 statistic and the Q statistic of a new batch with the confidence limits of the calibrated model, while fault identification and diagnosis is carried out by observing the contributions of the single process variables to these two statistics. The application of such methodology to uneven-length batches requires data alignment, consisting in equalization (all the variables need to be expressed at the same sampling rate across batches) and synchronization (all the landmarks for the variable trajectories need to be aligned across batches). Several methodologies for aligning uneven-length batches have been proposed, including dynamic time warping (Kassidas *et al.*, 1998) and its counterpart for synchronizing online data, relaxed greedy optimized warping (González-Martínez *et al.*, 2011). The adoption of these methodologies have some well-known downsides, such as a high computational cost with large datasets and the generation of artifacts when some batches are significantly shorter than the chosen reference (José M. González-Martínez *et al.*, 2014), and/or the starting and ending process conditions are different (González-Martínez *et al.*, 2011).

Methodologies based on feature extraction from the process variables time trajectories offer a workaround to batch data alignment. They are simple methodologies that are able to mathematically describe the problem retaining a relatively small dimensionality, however they do not describe a process in a time resolved manner. Furthermore, the calculation of the extracted feature must be performed when the whole time trajectory of a batch are available, their use in real time process monitoring is an open problem (Rendall *et al.*, 2019).

A challenge faced when using batch-wise unfolded models for monitoring of new batches is that only the current and previous measurement are available and measurements at future time points are missing. A workaround for this is the use of lagged multivariate models, however,

this workaround increase significantly the computational burden required for the modelling effort (Camacho *et al.*, 2008b).

In order to overcome these issues, the assumption-free modelling methodology (Westad *et al.*, 2015) was proposed. A number of assumptions required to apply standard Multiway-PCA to batch process monitoring are relaxed in this methodology: (*i*) it does not require batches alignment (neither equalization nor synchronization; (*ii*) it does not require batches to have the same starting and ending point; (*iii*) it does not require missing data imputation when doing online monitoring.

The methodology proposed by Westad *et al.* (2015) is composed of 2 main parts: the modelling step, where a common batch trajectory is modelled in the latent space of Multiway-PCA from a common start to a common end in chemical/biological time, and the monitoring step, where a new batch is monitored comparing it to the common batch trajectory produced in the previous step.

In the modelling step the common batch trajectory is generated through the identification of an optimal grid subdividing the latent space in order to identify similar process conditions in each of the subdivisions of the grid. An average process operating condition is identified for each subdivision, and these discrete average process operating conditions are then interpolated to obtain a continuous common batch trajectory. In the monitoring step a new batch is projected on the latent space and its distance from the common trajectory is calculated.

Unfortunately, neither the optimization problem to be solved to find the optimal grid, nor an algorithm solving the problem are available in literature. Furthermore, in the original work by Westad *et al.* (2015) it is recommended to use a linear or nonlinear interpolant to generate the common batch trajectory, however, no indications on when using an approach or the other are given. No indications are given as well regarding fault detection and fault identification and diagnosis through assumption-free monitoring. Lastly, no systematic comparison of assumption-free monitoring with a state-of-the-art monitoring methodology has been done to assess whether the methodology represent a real improvement with respect to previously available methodologies.

Although the assumption-free modelling methodology did not attract a wide interest in the scientific community (the original paper received 10 citations), it is known that this methodology is adopted in the industrial practice (Bano, 2023) and it is implemented in the commercial software Aspen Unscrambler.

In this study we propose guidelines for the practitioner to effectively implement the assumption-free monitoring methodology, filling in the gaps left in the original work. A mathematical formulation of the optimization problem that is to be solved to find the optimal grid for assumption-free modelling is proposed together with an algorithm solving it in a particular case well suited for practical applications. A comparison of the performance of two interpolant functions (linear and spline) for calculating the common batch trajectory is provided to

understand if higher degree interpolant functions have advantages with respect to a linear interpolation, which is the first choice to be made according to the parsimony principle. A method for carrying out fault detection together with a novel methodology for carrying out fault identification and diagnosis based upon a modified version of the Hotelling's T^2 contributions called relative T^2 contributions.

Lastly, a comparison between assumption-free monitoring and a state-of-the-art methodology based on Multiway-PCA is carried out on 5 case studies, a simulated SBR rubber polymerization process (Nomikos and MacGregor, 1994), an industrial LDPE batch polymerization process (Nomikos and MacGregor, 1995b), a simulated *Saccharomyces Cerevisiae* fermentation process (González-Martínez *et al.*, 2018), a simulated penicillin fermentation process (Birol *et al.*, 2002) and an industrial herbicide production process (García-Muñoz *et al.*, 2003).

6.2 Guidelines for the implementation of the assumption-free monitoring methodology

In this section, clear guidelines for the implementation of the assumption-free monitoring methodology will be given, analyzing each step of the algorithm described in Section 3.1.5.3.

6.2.1 Variable-wise unfolding

In this step, calibration data are unfolded in the variable direction, obtaining a 2-dimensional matrix that can be modelled through PCA.

6.2.2 Data preprocessing

In this step, the variable-wise unfolded data are preprocessed in the most appropriate way. In case the preprocessing technique is not specified, the authors recommend the practitioner to apply autoscaling to the data as it is the most commonly used approach for preprocessing data for latent variable modelling.

6.2.3 PCA modelling

In step 3 a PCA model is calibrated on the data unfolded in step 1 and preprocessed in step 2 (Wold *et al.*, 1998). In the original work by Westad *et al.* (2015) it is suggested to find the optimal number of PCs by cross-validation across batches. The authors of this study recommend limiting the maximum number of PCs to 2 when calibrating PCA for assumption-free modelling. The reasons behind this choice will be cleared in the next sections.

6.2.4 Grid optimization

Step 1 of the algorithm consists in finding a grid that can model the batch trajectory in the best way. The optimality criterion suggested in Westad *et al.* (2015) is to choose the grid that gives the most grid elements. A formulation of the optimization problem that needs to be solved to find the optimal grid is the following:

$$
\begin{aligned}
&\max_{\mathbf{n_c}} \mathrm{n}(\mathrm{V_c}) \\
&\text{s. t.} \\
&\mathrm{V_c} = \left\{ \mathrm{G}_{\boldsymbol{\phi}}, 1 \le \phi_i \le n_{c,i}, i < A \middle| \exists \mathbf{t}^{\mathrm{T}} \in \mathrm{G}_{\boldsymbol{\phi}} \cap \mathrm{b} \; \forall \; \mathrm{b} \in \mathrm{B}_1 \right\} \\
&\mathrm{G} = \prod_{ii=1}^{A} \left[g_{\mathrm{l},ii}, g_{\mathrm{u},ii} \right] \\
&\mathrm{b}_n = \{ \mathbf{t}^{\mathrm{T}} | \mathbf{t}^{\mathrm{T}} \in \text{calibration batch } n \} \\
&\mathrm{B} = \{ \mathrm{b}_1, \dots, \mathrm{b}_n \} \\
&\mathrm{B}_1 \subseteq \mathrm{B} \\
&\mathrm{n}(\mathrm{B}_1) = \beta \mathrm{n}(\mathrm{B}) \\
&\sum_{iii=1}^{\mathrm{n}(\mathrm{V_c})} \mathrm{n}\left(\left\{ \mathbf{t}^{\mathrm{T}} \in \mathrm{G}_{\boldsymbol{\phi},iii}, \mathrm{G}_{\boldsymbol{\phi},iii} \in \mathrm{V_c} \right\} \right) \ge \alpha N \\
&g_{\mathrm{l},j} \le \min\left(\mathbf{t}_j \right), \quad 1 \le j \le A \\
&g_{\mathrm{u},k} \ge \max(\mathbf{t}_k), \quad 1 \le k \le A \\
&0 < \beta \le 1, \quad 0 < \alpha \le 1 \quad ,
\end{aligned}
\tag{6.1}
$$

where $\mathrm{n}(*)$ represents the number of elements of a set, $\mathrm{V_c}$ is the set of valid cells, $\mathbf{n_c}[A \times 1]$ is the vector containing the number of subdivisions of grid G along each PC, $\mathrm{G}_{\boldsymbol{\phi}}$ is the cell of grid G at position $\boldsymbol{\phi}$, b_n is the set of rows $\mathbf{t}^{\mathrm{T}}$of matrix $\mathbf{T}$ belonging to calibration batch n, B is the set of calibration batches, B_1 is a subset of B containing a fraction β of the number of the elements of set B, α is a fraction of the number of rows $\mathbf{t}^{\mathrm{T}}$ of matrix $\mathbf{T}$ that must be included in the valid cells $\mathrm{G}_{\boldsymbol{\phi}}$. Some box constraints are defined, the lower and upper boundaries of the hyperrectangle G must be respectively lower than the minimum score and larger than the maximum score over each PC. Lastly, both α and β are bound between 0 and 1.

Westad *et al.* (2015) suggest to find the optimal grid performing an exhaustive search through a grid-search algorithm, in this study we therefore propose an algorithm that solves the optimization problem proposed in (6.1) for 2 PCs, shown in Figure 6.1:

1. choose the value of parameter α;
2. choose the maximum number of cells to be searched for each PC, $M_{c,1}$ and $M_{c,2}$, respectively;
3. find minimum and maximum calibration score value for each PC;

4. choose grid boundaries $\mathbf{g}_l$ and $\mathbf{g}_u$ respecting the constraints of (6.1) according to what has been found in Step 3;
5. assign $n_{c,1} = 1$;
6. assign $n_{c,2} = 1$;
7. build the grid with $\mathbf{n_c}$ subdivisions and calculate $n(V_c)$;
8. if $n_{c,2}$ is smaller than $M_{c,2}$, increment $n_{c,2}$ by 1 and return to step 7, else go to step 9;
9. if $n_{c,1}$ is smaller than $M_{c,1}$ increment $n_{c,1}$ by 1 and return to step 6, else go to step 10;
10. identify the combination $n_{c,1}$, $n_{c,2}$ among the one searched with the maximum number of valid cells and at least a fraction α of scores included in valid cells.

The boundaries of G, as well as α and β are user defined.

It must be noted that using a grid-search algorithm for solving (6.1) for a larger number of PCs produces a combinatorial explosion. In fact, the number of combinations tested by the grid-search algorithm increase exponentially with the number of PCs (e.g., to find the optimal subdivision searching from 1 to 10 cells for each of A PCs, 10^A combinations must be assessed) making the grid-search algorithm become rapidly very computationally expensive.

6.2.5 Calculation of batch mean value for each valid cell

In step 5 of the algorithm described in Section 3.1.5.3, the mean of all the scores included in each valid cell found in step 4 is calculated, furthermore, the overall mean of the scores of each one of the batches included in a valid cell is calculated.

6.2.6 Calculation of the common batch trajectory

In step 6 the overall means of the scores of all valid cell calculated in step 5 are interpolated to build the common batch trajectory Θ. This trajectory in the latent space represents the average conditions of the NOC batches used for PCA model calibration. The use of either a linear interpolation or a spline interpolation are suggested in the original work (Westad *et al.*, 2015), however, the authors of this study suggest to use a linear interpolation for building the common batch trajectory as no benefits from using more complex interpolating functions, such as spline, have been observed in the application to the case studies presented in Appendix A.

6.2.7 Calculation of confidence interval around the common batch trajectory

In step 7 the confidence intervals for T^2 and Q are calculated around the common batch trajectory. In all valid cells, the single-batch means calculated in step 6 are orthogonally projected onto the common batch trajectory and the distance of the mean of a batch in a cell from the common batch trajectory is assumed to be the length of the segment between the score and its projection. In case it is not possible to calculate the orthogonal projection of a score on

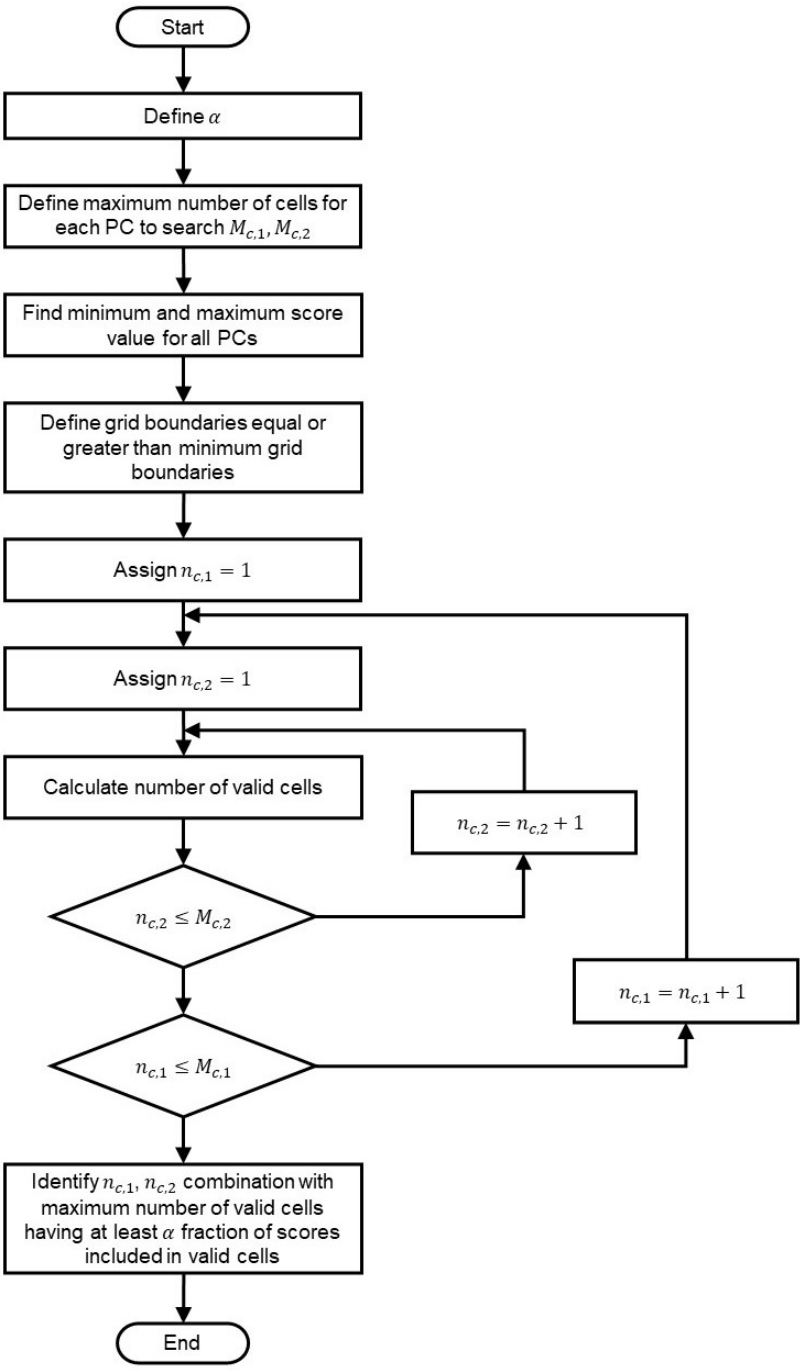


Figure 6.1: *Proposed grid-search based algorithm for finding the optimal grid subdivision in assumption-free modelling*

a segment, as shown in Figure 6.2, the distance from the score to the overall score mean in that valid cell is assumed to be the distance of the score from the common batch trajectory.
Performing these operations for all available batches in a certain valid cell results in a distribution of distances from the common batch trajectory.

6.2.8 Estimate the standard deviation around the common batch trajectory

The confidence interval is calculated as the value of the inverse normal cumulative distribution function with mean the mean of the distances of the batches mean from the common batch

(*a*) (*b*)

Figure 6.2: *(a) Calculation of the distance from the common batch trajectory (in blue) of a score (red circle) that cannot be orthogonally projected on the common batch trajectory). (b) Calculation of the distance from the common batch trajectory (in blue) of a score (red circle) that can be orthogonally projected on the common batch trajectory).*

trajectory, standard deviation the standard deviation of the batches mean from the common batch trajectory and the desired significance.

6.2.9 Calculate residual distance and its confidence limit for each valid cell

The value of the Q statistic for each sample belonging to a certain cell is calculated according to (3.7). The confidence limit for each valid cell is then calculated applying (3.19) to the values of Q of the samples belonging to the cell.

6.2.10 Calculate relative time for each sample in the calibration dataset

In order to calculate the relative time r_t for each sample in the calibration dataset, a large number of points (the authors suggest to consider a number of points in the order of 10^4) is sampled from the common batch trajectory calculated in Section 6.2.6. These points are progressively numbered according to the time progression of the common batch trajectory. For each sample in the calibration dataset, the closest point among the set of sampled points of the common batch trajectory is found. The relative time of that sample will be the fraction (or the percentage) between the number of that point on the common batch trajectory and the total number of points sampled from the common batch trajectory.

As the relative time is a time measurement, the authors of this work suggest to constraint it to be an increasing, monotonic function along a single batch.

6.2.11 Scale and center new observations

New observations are scaled and centered with respect to the calibration dataset when available.

6.2.12 Project new observation in the PCA model

After centering and scaling, in step 12 the new observation is projected on the PCA model calibrated in step 3 through (3.9).

6.2.13 Project new observation onto the common batch trajectory and calculate the distance from the common batch trajectory and the model

The score obtained by projecting the scaled and centered sample calculated in step 12 is then projected onto the common batch trajectory with the procedure described in Section 6.2.7. Its SPE statistic value is then calculated according to (3.7).

6.3 Fault detection and diagnosis using the assumption-free monitoring technique

In the algorithm described in the original work by Westad *et al.* (2015) it is not described how fault detection and diagnosis is carried out using the assumption-free monitoring technique as the technique is essentially presented as a visual methodology, however, it is known that it is industrially used for fault detection and diagnosis (Bano, 2023).
In this paragraph we propose a methodology for fault detection and diagnosis using the assumption-free monitoring technique.

6.3.1 Fault detection

The authors of this work suggest to use the Q statistic and the distance from the common batch trajectory confidence limits respectively for fault detection purposes.
If a consecutive number of samples in a batch violates either the Q or the common batch trajectory distance confidence limit, an alarm is triggered and the batch is considered faulty. Adopting the jargon of classification models, the consecutive number of samples must be tuned with the target of minimizing false positives, i.e., the number of normal batches classified as faulty (Magán-Carrión *et al.*, 2013; Rato *et al.*, 2016; Sanchez-Fernández *et al.*, 2015).

6.3.2 Relative contribution plots: fault diagnosis

In the assumption-free monitoring technique, fault detection is carried out through a dynamic control limit calculated along the common batch trajectory and through Q, however, for fault diagnosis purposes, when the fault is detected through the dynamic control limit, it is suggested to use the standard Hotelling's T^2 contributions (Westad, 2020).

When a fault is detected as a deviation from the common batch trajectory (and not as an excessive distance from the model plane, through Q), that is a deviation from the average conditions at a certain relative time, fault diagnosis is instead conducted by inspecting which variables contribute more to the deviation of the conditions of the faulty batch from the overall average condition (i.e., the origin of the axes of the latent space) over batches and over time. This approach is theoretically incorrect as the operating conditions of batch processes are inherently dynamic, therefore comparing the conditions of a batch at a certain relative time with the operating conditions averaged over the batch trajectories have no meaning. Furthermore, the use of the standard Hotelling's T^2 contributions in the assumption-free modelling technique brings to incoherent results (different fault diagnosis for the same fault) as well as completely wrong results, as will be shown in the results section.

For this reason, we propose the use of novel contributions called the relative T^2 contributions. The original variables for each point from the common batch trajectory Θ can be reconstructed as

$$\mathbf{X}_{\mathrm{ct}} = \Theta \mathbf{P}^{\mathrm{T}} \cdot \quad , \tag{6.2}$$

for each point of the common trajectory, its T^2 contributions $\mathbf{t}^2_{\mathrm{cont,ct}}{}^{\mathrm{T}}$, is calculated according to (6.2). When monitoring a new batch, for each sample its T^2 contribution, is calculated once again through (6.2), while its relative time r_t is calculated as described in Section 6.2.10. The relative T^2 contributions is calculated as the difference between the T^2 contributions of the new sample, $\mathbf{t}^2_{\mathrm{cont,new}}{}^{\mathrm{T}}$ and the T^2 contributions of the common batch trajectory point with the same relative time:

$$\mathbf{t}^2_{\mathrm{cont,rel}}{}^{\mathrm{T}} = \mathbf{t}^2_{\mathrm{cont,new}}{}^{\mathrm{T}} - \mathbf{t}^2_{\mathrm{cont,ct}}{}^{\mathrm{T}} \quad . \tag{6.3}$$

The relative T^2 contributions as calculated in (6.3) allow to assess which variable contributes more to the observed multivariate difference between the new batch conditions and the average conditions at the same relative time.

6.4 Case studies

In this study the assumption-free monitoring methodology is tested on 5 different case studies and its monitoring performance is compared with the performance of a state-of-the-art

methodology, batch-wise unfolding multiway PCA. In Table 6.1 the case studies characteristics are enlisted.

Table 6.1. *Case studies characteristics*

No.	Name	Process type	Operating mode	Industrial	Equalized data	Synchronized data	Faulty batches
1	SBR rubber emulsion copolymerization	Chemical	Fed-batch	No	Yes	Yes	2
2	LDPE polymerization	Chemical	Batch	Yes	Yes	Yes	1
3	*Saccharomyces Cerevisiae* fermentation	Biological	Batch	No	No	No	40
4	Penicillin manufacturing	Biological	Fed-batch	No	Yes	No	30
5	Herbicide crystals drying	Physical	Batch	Yes	No	No	38

The case studies have been selected for representing a wide range of batch and fed-batch processes that may be improved by implementing a data-driven monitoring methodology. Furthermore, the processes have a very different range of process duration both in terms of available measurement samples and in terms of time duration of the batches.

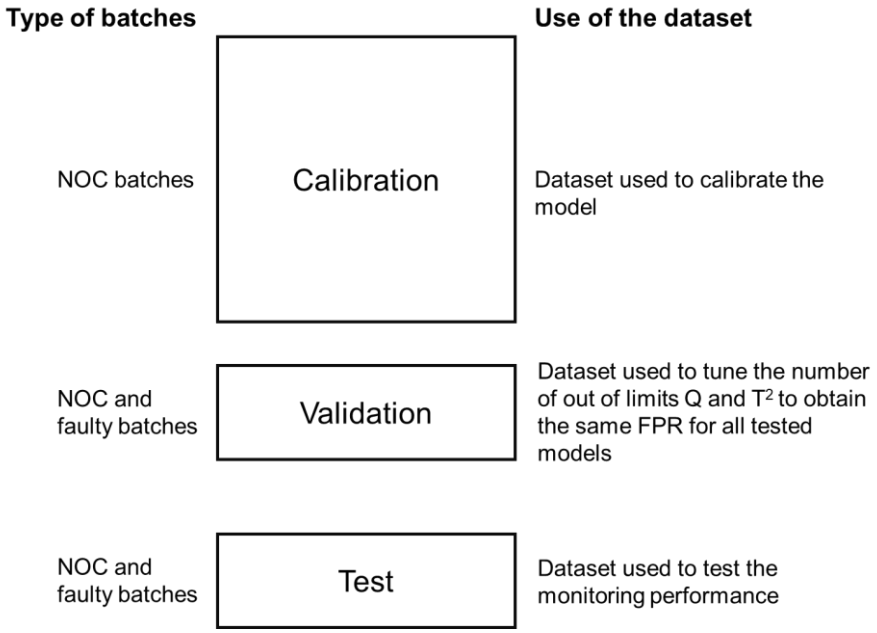


Figure 6.3 *Split of the batches into 3 datasets for comparing the assumption free methodology with the BWU MPCA methodology carried out in case studies 3,4 and 5; in case studies 1 and 2 the validation dataset was used also as test dataset.*

In case studies 3, 4 and 5, due to the sufficiently large number of batches available, the batches were split into 3 datasets: a calibration dataset where the monitoring models are calibrated, a validation dataset where the tunable parameters, namely the number of observations out of Hotelling's T^2 statistic confidence limit and the number of observations out of Q statistic confidence limit to trigger an alarm are tuned, and finally an external test dataset where the

monitoring performance of the monitoring models are assessed. In Figure 6.3 the datasets obtained splitting the batches with the abovementioned approach is shown.

6.4.1 Case study #1: simulated semibatch styrene/butadiene rubber (SBR) emulsion copolymerization

The process under investigation is the emulsion copolymerization of a styrene/butadiene copolymer to make a latex rubber for which a detailed first principle model has been developed (Broadhead *et al.*, 1985). The reactor is initially charged with seed SBR particles, initiator (S_2O_8), chain transfer agent (an aliphatic mercaptan), emulsifier (fatty acid soap), water and a small amount of styrene and butadiene monomers. Styrene and butadiene are fed to the batch at a constant rate for the remainder of the batch duration. The geometry of the reactor is assumed to be cylindrical, and the reactor is assumed perfectly mixed. The temperature inside the reactor is controlled by manipulating the cooling water flowrate of the jacket.

The reaction starts with the radical decomposition of S_2O_8:

$$S_2O_8 \rightarrow 2SO_4^- \cdot \quad , \tag{6.4}$$

$$SO_4^- \cdot + M \rightarrow SO_4^- + R \cdot \quad , \tag{6.5}$$

where R can be either styrene (St) or butadiene (Bu).

The propagation reactions can happen in the presence of radical monomers:

$$\sim St \cdot + St \rightarrow \sim StSt \cdot \quad , \tag{6.6}$$

$$\sim St \cdot + Bu \rightarrow \sim StBu \cdot \quad , \tag{6.7}$$

$$\sim Bu \cdot + St \rightarrow \sim BuSt \cdot \quad , \tag{6.8}$$

$$\sim Bu \cdot + Bu \rightarrow \sim BuBu \cdot \quad . \tag{6.9}$$

Radical termination is assumed to occur only in the polymer phase because of chain transfer to monomer, polymer or modifier.

The set of online measured process variables for this case study is shown in Table 6.2. The unit of measurement of variables 2 and 3 are not reported for confidentiality reasons.

Variability in the dataset is obtained by adding noise to the initial charge purity and butadiene flowrate, as well as in the feed temperature measurements (Nomikos and MacGregor, 1994).

Table 6.2. *Case study #1: variables measured in real time.*

Variable no.	Variable name	Units of measurement
1	Time	min
2	Styrene flowrate	
3	Butadiene flowrate	
4	Feed temperature	°C
5	Reactor temperature	°C
6	Cooling water temperature	°C
7	Reactor jacket temperature	°C
8	Latex density	g/L
9	Total conversion	[-]
10	Net energy released	J/min

Batch duration is 1000 min and a measurement for the first 8 process variables is available once every 5 minutes, while variable 9 and 10 are estimated online from the energy balance around the reactor.
The calibration dataset is constituted of 45 NOC batches while the validation dataset is constituted of 8 batches, 6 NOC batches and 2 faulty batches: one having a 30% larger impurity contamination in butadiene feed from the beginning of the batch, the other having a 50% larger impurity contamination at 500 minutes from the beginning of the batch.

6.4.2 Case study #2: industrial low-density polyethylene (LDPE) batch polymerization

In this case study an industrial (DuPont) LDPE polymerization process has been considered (Nomikos and MacGregor, 1995a). The recipe is constituted of two processing steps, the batch duration is two hours. In the first step, reactants and solvent are loaded into the reactor and the correct rate of pressure and temperature changes are established. The solvent through which the reactants are loaded in the reactor is then vaporized and removed from the reactor. Due to the large vaporization rate the reactor is not stirred. After an hour the first processing step ends, and the second step starts. During this processing step the polymerization reaction is completed. After that, the product is discharged to downstream processing.
In Table 6.3 are enlisted the online measured process variables. Units of measurement are unavailable due to confidentiality reasons.

Table 6.3. *Case study #2: variables measured in real time.*

Variable no.	Variable name
1	Temperature 1
2	Temperature 2
3	Temperature 3
4	Pressure 1
5	Flowrate 1
6	Temperature 1 heat/cool
7	Temperature 2 heat/cool
8	Pressure 2
9	Pressure 3
10	Flowrate 2

100 measurements carried out at regular interval along the duration of each batch are available. The calibration dataset is constituted of 50 NOC batches, while the validation dataset is constituted of 5 batches, 4 NOC batches and 1 presenting a fault from the beginning of the batch.

6.4.3 Case study #3: simulated batch manufacturing of Saccharomyces Cerevisiae

We consider a batch fermentation process that manufactures *Saccharomyces Cerevisiae*. A simulator (González-Martínez *et al.*, 2018) which implemented a model for the aerobic growth of *Saccharomyces Cerevisiae* on glucose limited medium is used. A detailed description of the model can be found in the work of Lei *et al.* (2001). According to the model, fermentation is carried out in four different steps: (i) lag phase, (ii) first exponential growth phase, (iii) second exponential growth phase and (iv) stationary phase. In (i), the microorganism adapts to the culture media before starting the reproduction process; this phase usually lasts around 2 hours. In (ii) cells are not able to consume the whole amount of glucose present in the medium, hence ethanol is produced, and pyruvate and acetate are excreted. At the end of (ii) glucose is completely consumed by the growing cells and in (iii) cells start to grow consuming ethanol and producing acetate.

In Table 6.4 the list of real time measured variables for this case study is made available.

Each batch lasts 35 h in absolute time but the available data are not equalized, hence both the number of samples per batch and the sampling rate vary. Variability is added to the initial conditions as gaussian noise with a standard deviation equal to 10% of each initial condition value, furthermore, low magnitude additive measurement noise has been added to the time trajectories of the online measured variables (González-Martínez *et al.*, 2018).

Table 6.4. *Case study #3: variables measured in real time.*

Variable no.	Variable name	Units of measurement
1	Glucose concentration	g/l
2	Pyruvate concentration	g/l
3	Acetaldehyde concentration	g/l
4	Acetate concentration	g/l
5	Ethanol concentration	g/l
6	Biomass concentration	g/l
7	Active cell material	[-]
8	Acetaldehyde dehydrogenase	[-]
9	Specific oxygen uptake rate	mmol/(g h)
10	Specific CO2 evolution rate	mmol/(g h)
11	Simulation time	h

The calibration dataset is constituted of 40 NOC batches. The mean number of measurements for calibration batches is 211, with a standard deviation of 32 measurements.

Two different types of faults are introduced in validation and test batches at different moments in time and with different magnitudes.

The first type of fault is generated by modifying an internal rate constant associated with the reaction describing the glucose uptake system and the glycolytic pathway. This fault does not relate to an abnormal behaviour related to specific biochemical changes in the metabolic

network, but to an upset in operating conditions that may alter the model kinetics. The second type of fault is a sensor fault that introduces a bias in the biomass concentration sensor.
The validation dataset is constituted of 3 NOC batches and 21 faulty batches, while the test dataset is constituted of 2 NOC batches and 18 faulty batches.

6.4.4 Case study #4: simulated fed-batch manufacturing of penicillin

A fed-batch manufacturing process producing penicillin is considered. The process is simulated using Pensim (Birol *et al.*, 2002). A simplified P&ID of the process is shown in Figure 6.4.
The recipe for penicillin manufacturing is based on two processing steps:

1. A batch culture step, where the reactor is initially loaded with Penicillium Chrysogenum and glucose from tank T4, and the reaction starts; this step ends when the concentration of glucose in reactor R2 drops below an assigned threshold;
2. A fed-batch step, where pH is automatically controlled through the addition of acid from tank T2 and base from tank T3; during this step, glucose and air are fed constant rates. The end-point condition is reached when the total volume of glucose fed to R2 during this step reaches 14 L (Sun *et al.*, 2011).

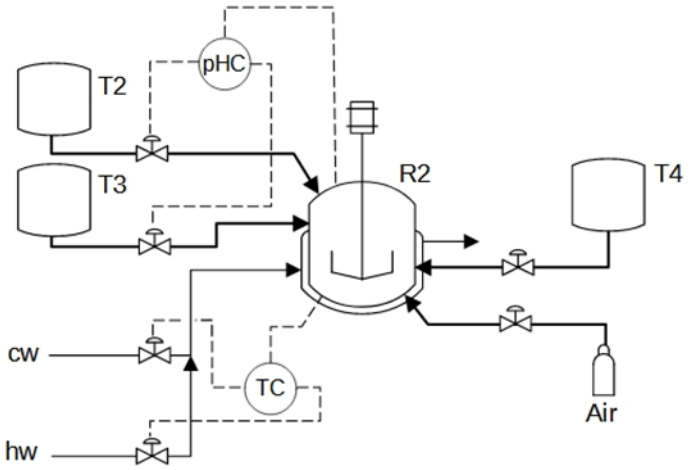


Figure 6.4 *Case study #4: simplified piping and instrumentation of the simulated process for penicillin manufacturing*

Real time measurements of some process variables are available as listed in Table 6.5. Measurement noise is simulated in the form of additive random numbers sampled from a normal distribution with zero mean and standard deviation $stdev$ as indicated in Table 6.5. Process variability is generated by randomly changing the values of some initial conditions and some operating variables, as detailed in Table 6.6. Further variability is generated by assuming that the threshold glucose concentration determining the switch between operating steps 1 and 2 randomly varies between 0.3 and 7 g/L.

Table 6.5. *Case study #4: variables measured in real time.* $stdev$ *is the standard deviation of a zero-mean normal distribution of random numbers.*

Variable no.	Variable name	Units	$stdev$
1	Time	h	0
2	Aeration rate	L/h	0.083
3	Agitator power	W	0.167
4	Glucose feed rate	L/h	0.00083
5	Glucose feed temperature	K	0.167
6	Glucose concentration	g/L	0
7	Dissolved O_2	g/L	0.0067
8	Biomass concentration	g/L	0
9	Penicillin concentration	g/L	0
10	Bulk volume	L	0.033
11	Dissolved CO_2	mmol/L	0
12	pH	[-]	0.0167
13	Fermentor temperature	K	0.167
14	Generated heat	cal	0
15	Acid flow rate	L/h	$3.3 \cdot 10^{-7}$
16	Base flow rate	L/h	$3.3 \cdot 10^{-6}$
17	Cooling/heating water flowrate	L/h	0.83
18	Cumulated acid flow rate	L	0
19	Cumulated base flow rate	L	0
20	Cumulated glucose feed rate	L	0

The measurements for the online variables are available once every 15 min. The calibration dataset is constituted of 30 NOC batches with mean batch length of 200 h (that translates to a mean of 800 samples per batch).

Two different types of faults are introduced in validation and test batches at different moments in time and with different magnitudes.

Table 6.6. *Case study #4: nominal initial conditions, nominal operating variables, and variability around them (*ϵ *is sampled from a standard normal distribution).*

Initial condition	Units	Nominal value
Glucose concentration	g/L	$15 + \epsilon$
Dissolved oxygen	%	1.16
Biomass concentration	g/L	0.1
Penicillin concentration	g/L	0
Culture volume	L	$150 + 10\epsilon$
CO_2 concentration	mmol/L	$0.75 + 0.05\epsilon$
Hydrogen ion concentration	mol/L	$10^{-5+0.1\epsilon}$
Fermentor temperature	K	298
Generated heat	kcal/h	0

Operating variable	Units	Nominal value
Aeration rate	L/h	8
Agitator power	W	$30 + \epsilon$
Glucose feed rate	L/h	$0.04 + 0.0025\epsilon$
Glucose feed temperature	K	296
Culture volume	L	$150 + 10\epsilon$
pH	[-]	5
Fermentor temperature	K	298

The introduced faults are sensor faults of the aeration rate sensor and of the glucose feed rate sensor happening in different moments and at different magnitudes in faulty batches.
The validation dataset is constituted of 7 NOC batches and 20 faulty batches, while the test dataset is constituted of 2 NOC batches and 10 faulty batches.

6.4.5 Case study #5: industrial batch drying for herbicide manufacturing

An industrial batch drying process for the production of a herbicide is considered in this case study. The recipe for the batch drying process has 4 main steps (García-Muñoz *et al.*, 2003):

1. The batch is loaded with a wet cake with variable mass and unknown solvent content;
2. the dryer is heated with hot water and the agitator is activated at low speed, a temperature increase is observed;
3. when the required operating conditions are reached, the agitator speed is switched up and the temperature increase becomes faster until the maximum temperature is reached: at this point the agitator speed is turned down again;
4. after the temperature peak is reached, the dried product is cooled down while the agitator remains active until the product discharge.

The vaporized solvent is condensed and recovered in a separated tank.
In Table 6.7 the real time measured variables are enlisted. Units of measurement are not available due to confidentiality reasons.
The dataset was made available by the FMC corporation.
The calibration dataset is constituted of 28 NOC batches with a mean of 117 samples available per batch.

Table 6.7. *Case study #5: variables measured in real time.*

Variable no.	Variable name
1	Tank level
2	Differential pressure
3	Dryer pressure
4	Power
5	Agitator speed
6	Torque
7	Jacket temperature setpoint
8	Jacket temperature
9	Dryer temperature setpoint
10	Dryer temperature

In the validation and test datasets, two different types of faulty batch are present. One type of faulty batch has an off-spec product at the end of batch, while the other type is off-spec but it exhibit an anomalously large amount of solvent at the end of batch.
The validation dataset is constituted of 2 NOC batches and 29 faulty batches, while the test dataset is constituted of 1 NOC batch and 9 faulty batches.

6.5 Results

In this paragraph the assumption-free monitoring technique will be applied according to the proposed guidelines and its monitoring performance will be compared with the BWU MPCA monitoring technique. Fault detection using the BWU MPCA methodology is carried out using the Q and T^2 statistics, although also a different methodology using the SPE statistic can be used for real time monitoring (Nomikos and MacGregor, 1995a). The assumption-free monitoring technique will be calibrated using both a linear and a spline interpolant and their monitoring performance will be compared. The monitoring techniques performance will be compared both through metrics of detection strength (true positive rate, TPR; false positive rate, FPR; Rato *et al.*, 2016) and detection speed (average run length, ARL; Rato *et al.*, 2018). Furthermore, the fault diagnosis will be carried out through the relative contributions plot and it will be compared with the standard T^2 contributions plot.
The results from the 5 case studies, summarised in Table 6.8, are discussed in detail in the remainder of this section.

Table 6.8: *Monitoring performance comparison between assumption-free monitoring and BWU MPCA*

Case study	Average batch length	Methodology	PC no.	o.o.l. Q	o.o.l. T^2	TPR	FPR	ARL[samples]
1	200	**Assumption-free (linear)**	2	5	48	100	16.7	86
		Assumption-free (spline)	2	5	60	100	16.7	96
		BWUMPCA	3	3	2	100	16.7	85
2	100	**Assumption-free (linear)**	2	4	15	100	0	6
		Assumption-free (spline)	2	4	15	100	0	6
		BWUMPCA	3	3	1	100	0	21
3	211	**Assumption-free (linear)**	2	9	35	100	0	48
		Assumption-free (spline)	2	9	35	100	0	49
		BWUMPCA	3	300	3	50	0	83
4	750	**Assumption-free (linear)**	2	16	22	96.5	10	196
		Assumption-free (spline)	2	16	22	96.5	10	231
		BWUMPCA	3	700	3	62.1	30	400
5	115	**Assumption-free (linear)**	2	2	24	94.7	10	68
		Assumption-free (spline)	2	2	35	86.8	10	71
		BWUMPCA	3	21	5	60.5	33.3	94

6.5.1 Results for case study #1

The calibration dataset is used to calibrate a BWU MPCA model with 3 PCs ($R^2 = 30.6\%$) and an assumption-free model with 2 PCs ($R^2 = 56.9\%$). In Figure 6.5 it is shown that a 6×7 grid is found by assumption-free model calibration with 11 valid cells and a percentage of scores included in valid cells of 96.6%. In the score plot a concentration of points around the origin of the axes is observed (cells 8-10), while a much less dense scores are present in the initial part of the batch (cells 1-7).

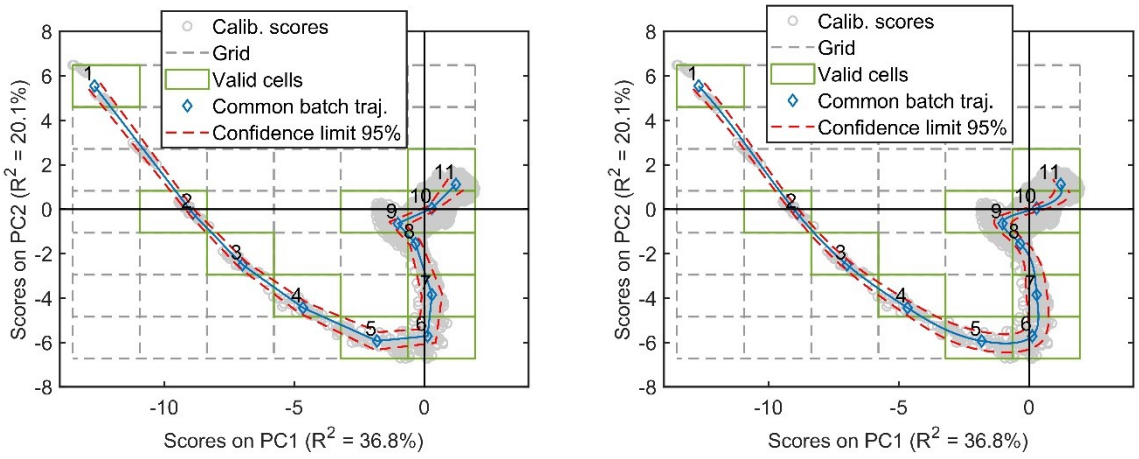


Figure 6.5: *(a)Score plot with assumption free (linear interpolant) common batch trajectory and dynamic control limit represented as a blue line and red dashed lines, respectively. (b) Score plot with assumption free (spline interpolant) common batch trajectory and dynamic control limit represented as a blue line and red dashed lines, respectively.*

In Table 6.8 the monitoring performance of the assumption-free models (both linear and spline) are compared with the performance of the BWU MPCA model.

In this case study, BWU MPCA performs better both in terms of detection strength, as all batches are correctly classified, and detection speed, as it requires less than half the time of the assumption-free methodology to detect faults.

The cause of the better performance shown by the BWUMPCA model lies in the process dynamics. In fact, in Figure 6.6 the time trajectories of 2 relevant variables for this case study are shown and it is observed that the dynamics is confined before sample 50, while in the remainder of the batch very small changes are observed in total conversion, while oscillations around 46.8°C are observed in cooling water temperature.

Because of this, a large distance in the scores corresponding to the initial samples of the profiles is observed, while all the scores after sample 50 are tightly grouped.

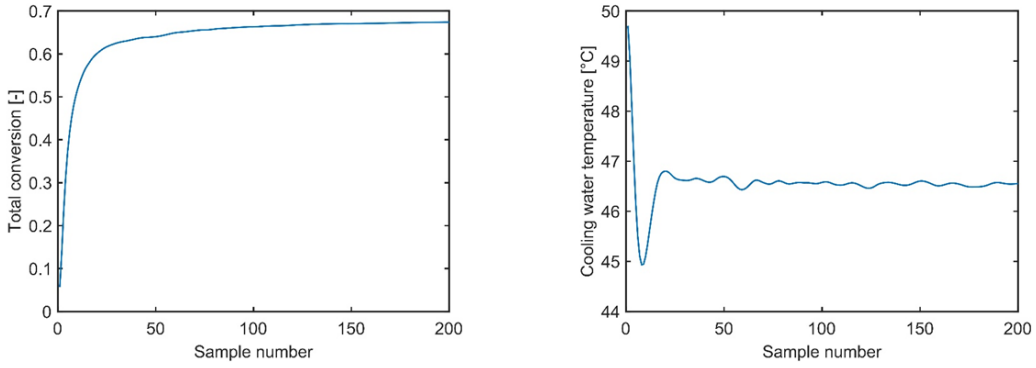


Figure 6.6: *(a) Time trajectory of variable 9 of batch 12 of case study #1. (b) Time trajectory of variable 6 of batch 12 of case study #1.*

6.5.2 Results for case study #2

A BWU MPCA model with 3 PCs ($R^2 = 64.5\%$) and an assumption-free model with 2 PCs ($R^2 = 85.5\%$) are calibrated.

A 10×5 grid is considered optimal, with 18 valid cells and a percentage of scores included in valid cells of 98.4%. The monitoring performance of these models in case study #2 are shown in Table 6.8. In this case study the performance of the assumption-free and BWU MPCA are equivalent in terms of detection strength, however the assumption-free model, on average detects a fault in a little less than 1/3 of the time required to BWU MPCA for the detection of a fault.

Furthermore, the assumption-free monitoring approach with a linear interpolant is preferred to the spline interpolant based approach as they have the exact same performance, but the model based on the linear interpolant is simpler.

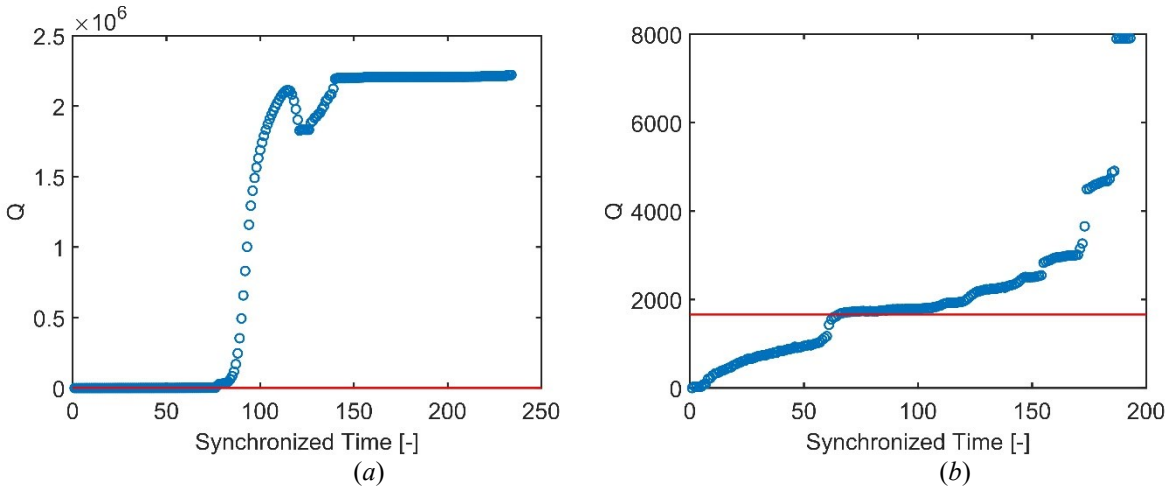
6.5.3 Results for case study #3

The calibration dataset of Case Study #3 has been used to calibrate a BWU MPCA model with 4 PCs ($R^2 = 39.9\%$) and an assumption-free model with 2 PCs ($R^2 = 68.9\%$). An 8×7 grid is found optimal for modelling the studied process, with 18 valid cells and a percentage of scores included in valid cells of 95.1%. The validation dataset has been used to tune the number of out of limits values of Q and T^2 and the performances of the models have been evaluated on an external test set, the results of the performance evaluation are reported in Table 6.8.

It is observed that for the BWU MPCA model, the number of out of limits samples on the Q statistic required for triggering an alarm is larger than the batch length after alignment (212 samples). This value was obtained by tuning the parameter in validation with the aim of minimizing false positive batches.

In Figure 6.7 it is shown that although the Q statistic of faulty batch in Figure 6.7a reaches much larger values, the NOC batch in Figure 6.7b goes past the control limit at around half the duration of the batch. This is a known problem of the use of the Q statistic in monitoring using a BWUMPCA model when the batch-wise unfolded dataset have a large number of variables (Reis *et al.*, 2021). This issue reduces drastically the monitoring performance of the BWU MPCA in this case study as it can only identify correctly as faulty the 50% of the test dataset batches. The assumption-free modelling, being based upon variable-wise unfolding does not suffer of issues in using the Q statistic for monitoring purposes, and the conjoint use of both Q and T^2 for monitoring produces much better performance, with a TPR reaching 100%.

Comparing the performance of the assumption-free monitoring technique using a linear or a spline interpolant, in Table 6.8 it is shown that they have similar performance in this case study, therefore a model using a linear interpolant is considered the best choice in compliance with the parsimony principle.



(*a*) (*b*)

Figure 6.7: *(a) Q statistic time trajectory during the monitoring of a faulty batch with BWUMPCA from the test dataset of case study #3. (b) Q statistic time trajectory during the monitoring of a NOC batch with BWU MPCA from the test dataset of case study #3.*

In Figure 6.8 the score plot with the monitoring scores of a faulty batch from the test dataset (test batch #31) is shown.

As shown in Figure 6.8, fault detection was successfully carried out as the scores of the faulty batch (the bordeaux crosses) are outside the dynamic control limit (the red dashed lines) for more than 30 samples, hence an alarm is triggered.

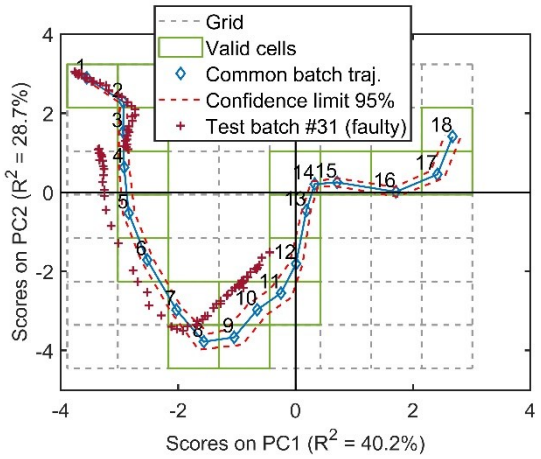
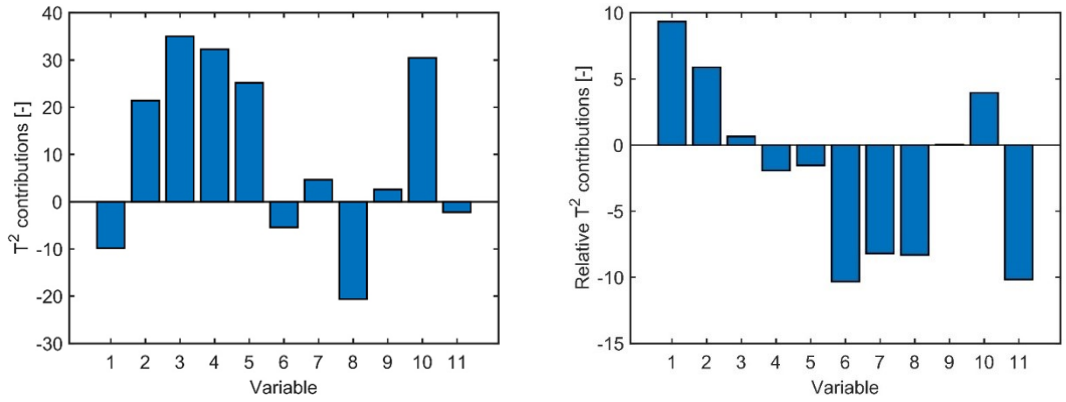


Figure 6.8: *Score plot showing the scores of test batch #31 calculated during the monitoring steps.*

The fault affecting the considered batch consists of the introduction of a -3 g/L bias in variable 6 (*biomass concentration*) from the beginning of the batch due to a faulty sensor.

The standard T^2 contribution plot is compared with the proposed relative T^2 contribution plot for the point triggering the alarm, respectively, in Figure 6.9a and 6.9b.



(*a*) (*b*)

Figure 6.9: *(a) T^2 contributions of the sample that triggered the alarm in test batch #31. (b) relative T^2 contributions of the sample that triggered the alarm in test batch #31.*

In Figure 6.9a it is shown that the most important variables identified by standard T^2 contributions are variable 2 to 5 (*Pyruvate, Acetaldehyde, acetate and Ethanol concentrations*), variable 8 (*Acetaldehyde dehydrogenase*) and variable 10 (*Specific CO_2 evolution rate*). Therefore, in this case, looking at the standard T^2 contribution plot produces a completely wrong fault diagnosis, while looking at Figure 6.9b it can be seen that the most important variable is variable 6, which is the main variable affected by the fault, hence, the proposed relative T^2 contributions are much more effective in this case than the standard T^2 contributions at diagnosing the fault.

The main reason for this difference between relative and standard T^2 contributions value is due to the direction that is considered in the score plot for calculating the contribution. In fact, the T^2 statistic (on which the standard T^2 contributions are calculated) used for calculating the standard contribution plots is a function of the norm of the vector joining the origin of the axes to the score that triggered the alarm in the currently monitored batch. Instead, the vector considered when calculating the relative T^2 contributions is the one joining the score that triggered the alarm in the currently monitored batch with the point at the same relative time in the common batch trajectory.

Proof of the previous statement is that in some lucky situations where the two vectors share the same direction the relative and the standard T^2 contribution plots work in the same way, as for the batch shown in Figure 6.10.

In this case, a fault affecting the kinetic constant of a reaction describing the microorganisms glucose uptake system and glycolitic pathway is simulated.

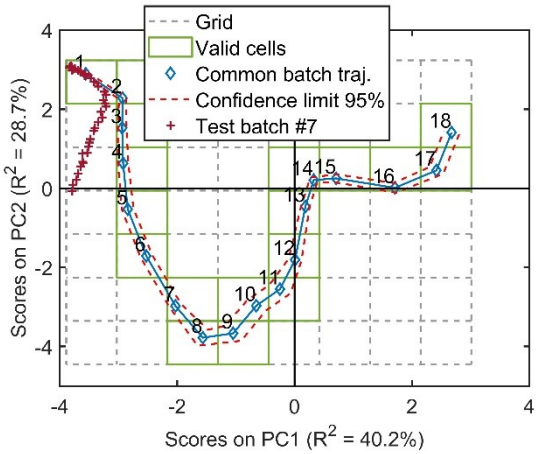
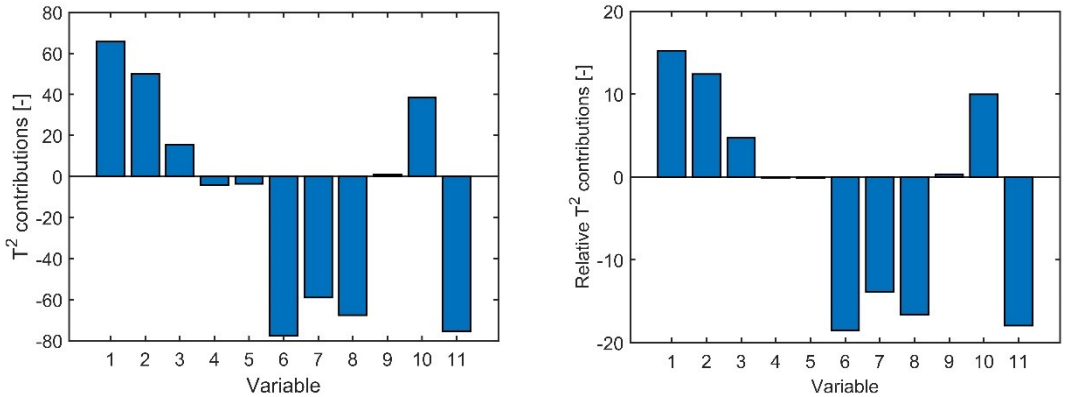


Figure 6.10: *Score plot showing the scores of test batch #7 calculated during the monitoring steps.*

As shown in Figure 6.10, the direction of the vector joining the origin of axes to the point of the test batch #7 that triggered the alarm (the red cross that lies onto the x-axis) and the direction of the vector joining the same point of test batch #7 and the common batch trajectory are almost parallel.

As a result of this situation, in Figure 6.11a and 6.11b it is observed that both the T^2 contribution plot and the relative T^2 contribution plot show the exact same result except for the absolute value, which is not relevant when using these plots (Miller *et al.*, 1998).

Both plots correctly show that *glucose concentration, pyruvate concentration, biomass concentration, active cell material and acetaldehyde dehydrogenase* (variables 1,2,6,7,8) are affected by the fault.



(*a*) (*b*)

Figure 6.11: *(a) T^2 contributions of the sample that triggered the alarm in test batch #7. (b) relative T^2 contributions of the sample that triggered the alarm in test batch #7.*

This last result shown asserts that the use of the T^2 contribution is risky in assumption-free modelling, as it can occasionally results in a correct fault diagnosis because of an incidental parallelism between two vectors, illuding the practitioner that this is the correct approach to fault diagnosis with this monitoring approach. The relative T^2 contribution plot instead yield a

correct fault diagnosis in all situation as it shows variables that contribute to a deviation of the monitored batch from the common batch trajectory by design.

6.5.4 Results for case study #4

In case study #4 a BWU MPCA model with 2 PCs ($R^2 = 42.3\%$) and an assumption-free model with 2 PCs ($R^2 = 58.0\%$) are calibrated onto the calibration dataset. A 10×3 assumption-free model grid was found to be optimal with 8 valid cells and 95.1% scores included in valid cells. The monitoring performance on the test dataset, together with the value of the parameters tuned on the validation dataset are shown in Table 1.

The number of out of limits Q statistic values for triggering an alarm set to 700 shows that also in this case study BWU MPCA suffers of the same issue affecting the Q statistic, resulting in a limited monitoring performance, whereas assumption-free monitoring performs much better both in terms of detection strength and detection speed, with around half the value of ARL with respect to BWU MPCA. The spline-based assumption-free monitoring approach results to have the same detection speed performance as the linear-based monitoring while being slightly slower than the model adopting a linear interpolant. For these reasons, an assumption-free monitoring approach with a linear interpolant is considered the best monitoring approach to be used for this case study.

In Figure 6.12 two monitored batches, test batch #24 and test batch #26, are shown. Both batches are affected by the same fault, a 30% increase in the glucose feed at half the batch length. The scores from these batches are represented from the beginning up to the sample that triggered the alarm.

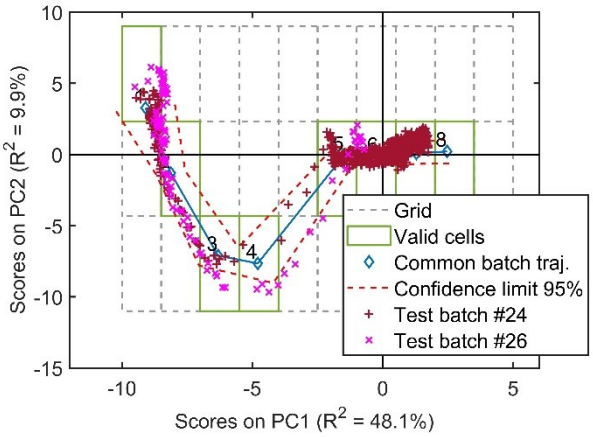


Figure 6.12: *Score plot for case study #4 with two faulty batches(same fault) projected: test batch #24 and test batch #26*

As Shown in Figure 6.12, test batch #26 have been detected much earlier than test batch #24, in fact the scores of test batch #26 are all in the left half plane, while the scores of test batch #24 go beyond the vertical axis into the first quadrant of the score plot.

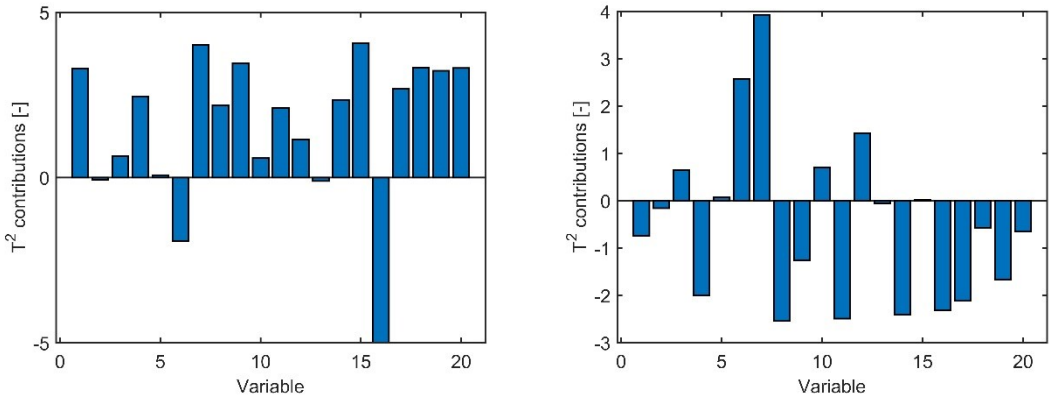


Figure 6.13: *(a) T^2 contributions of the sample that triggered the alarm in test batch #24. (b) T^2 contributions of the sample that triggered the alarm in test batch #26.*

In Figure 6.13a and 6.13b standard T^2 contributions for both batches are shown. In this case, for two batches with the same fault, a different fault diagnosis is obtained through the standard T^2 contribution plots, while, when using the relative T^2 contribution plots, as shown in Figure 6.14a and 6.14b, a coherent result is obtained indicating the same fault diagnosis for both batches.

This brings us to the conclusion that not only using the standard T^2 contributions results in a wrong fault diagnosis, as shown in case study #3, but the fault diagnosis results obtained through the T^2 contributions shows incoherences when dealing with different batches with the same fault, while relative T^2 contributions solve this problem as they results in the same fault diagnosis for batches with the same fault (except for the absolute value of each contribution).

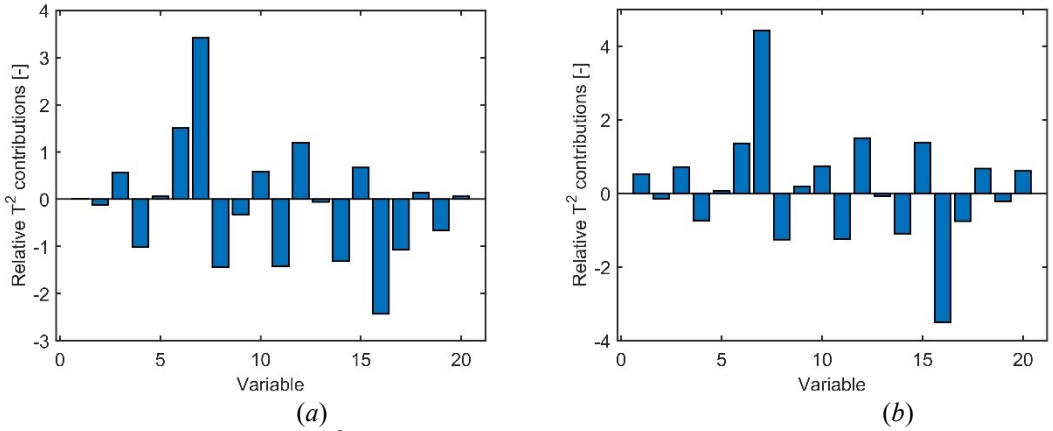


Figure 6.14: *(a) relative T^2 contributions of the sample that triggered the alarm in test batch #24. (b) relative T^2 contributions of the sample that triggered the alarm in test batch #26.*

6.5.5 Results for case study #5

The calibration step is carried out calibrating a 5 PCs BWU MPCA model ($R^2 = 63.2\%$) and a 2 PCs assumption-free model ($R^2 = 67.1\%$) with an optimal 2×3 grid, 4 valid cells and 98.2% of scores included in valid cells.

Results on the external test dataset are reported in Table 1.

In this case study the Q statistic value do not appear to be affected by the issues observed in case studies #3 and #4, as the tuned number of samples with out of limits Q statistic is much smaller than the batch length. This is due to the mean batch length (115 samples) being smaller than in the case studies presenting the problem, therefore obtaining a less fat batch-wise unfolded matrix.

Even though in this case study the conditions for the application of BWU MPCA appear to be more favourable than in the last two, in Table 5 it is shown that assumption-free model outperforms it also in this case both in terms of detection strength and detection speed.

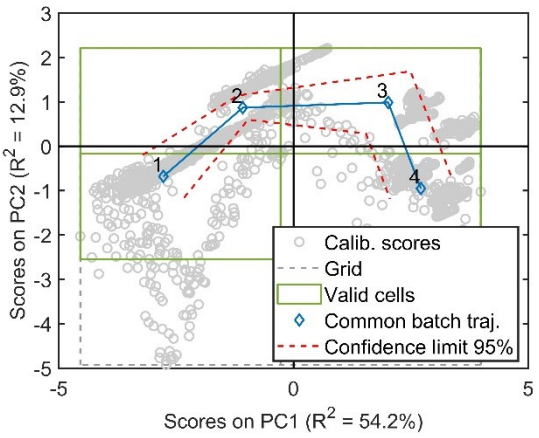


Figure 6.15: *Score plot for case study #5*

In Figure 6.15 it is shown the score plot for case study #5. The calibration dataset shows a very large variability, in fact a time trajectory for the historical batches is hardly (it not at all) visible in the score plot. In this situation the dynamic control limit (red dashed lines) is assumed to have a very poor performance, and in fact, in Table 5, a large number of samples with scores out of the dynamic control limit (o.o.l. T^2) are required to trigger an alarm using this statistic.

From further investigation resulted that all faulty batches are in fact detected by the assumption-free model through the Q statistic control limit.

Therefore, however being in a disadvantageous situation with this dataset and having the T^2 statistic not working for monitoring purposes, the dynamically defined Q control limit calculated through the assumption-free model is still much more efficient for monitoring purposes than the BWU MPCA model.

6.6 Conclusions

In this work, a set of guidelines for the implementation of the assumption-free monitoring technique were provided to simplify the implementation of the methodology for the practitioner. Light was shed on the optimization problem underlying the calibration of an assumption-free model and an algorithmic implementation to solve the problem in the most commonly implemented case was provided. Furthermore, clear indications for performing fault detection with an assumption-free model have been provided together with a new methodology for fault diagnosis called relative T^2 contribution plot. This novel methodology improves the standard T^2 contribution plot in terms of correctness of the fault diagnosis, as it is able to identify more effectively the variables causing a deviation from the common batch trajectory, and in terms of robustness of results, as it identifies consistently the same variables as the cause of the deviation from the common batch trajectory in different batches having the same fault.

The assumption-free monitoring implemented according to the proposed guidelines was tested on 5 datasets, 2 with even-length and 3 with uneven-length batches, coming from different industrial sectors and processes of different nature (chemical, biological and physical). The performance of the methodology was compared with the performance of BWU MPCA, a state-of-the-art method for batch process monitoring.

The assumption-free methodology consistently outperformed BWU MPCA (although the number of principal components of the underlying MPCA model is limited to 2), especially in cases when the number of time samples of the batches is large, as it is not affected by the deterioration of the Q statistic monitoring performance observed in BWU MPCA. Furthermore, the definition of a Q statistic control limit varying with relative time allows it to outperform BWU MPCA even in situations where a large variability in the calibration batches degradates the performance of the dynamic control limit.

The only situation in which the model performance is not better than the BWU MPCA performance is when the dynamics of the studied process is compressed in a very small fraction of the variables time trajectories.

The assumption-free monitoring methodology is also less computationally expensive than BWU MPCA when batch alignment must be carried out, as this technique does not require such a preprocessing step, and instead it implicitly performs a batch alignment building the relative time during the common batch trajectory construction.

Further improvements to the proposed guidelines regards:

- Generalize the algorithm solving the grid optimization problem to an arbitrary number of PCs, introducing heuristics in order to avoid the combinatorial explosion obtain applying grid searching to n parameters;
- building a grid optimized with respect to the density of the scores in the score plot in order to describe more efficiently the process variability;
- generalizing the grid to an optimal mesh with arbitrary geometry.

Conclusions and future perspectives

Batch processing is ubiquitous in several industrial sectors manufacturing high value-added products due to its flexibility in set up and operations. Furthermore, it is relatively simple to design even if only partial knowledge of the underlying mechanism of the process is available. However, its flexibility allows for a large amount of variability to enter the process, making the task of running the process in controlled conditions and producing a product of consistently high quality much harder. This task is made even harder because the product quality is usually assessed only at the end of a batch, due to the time required by laboratory assays and to reduce the workload on the quality control laboratory. Furthermore, the variability entering the process can reduce the efficiency of the process in terms of utility consumption as well as increasing its cycle time, making the production scheduling more complex. The abovementioned characteristics of batch processes render the development and the implementation of process monitoring techniques paramount to run batch processes efficiently. The large number of process variables observed and recorded every few seconds in a modern chemical plant, available thanks to the technological advancements triggered by the Industry 4.0 initiative, may lead human supervisors to an overload of information if observed separately. Furthermore, observing the process variables one-at-a-time may return insufficient, incomplete and unreliable information. Multivariate statistical methods allow to reduce the dimensionality of the process monitoring problem to a latent subspace, dealing with spatial and serial correlation, multicollinearity, noise and missing data at the same time.

The objective of this Dissertation was the practical implementation of Industry 4.0 methodologies for monitoring the performance of batch processes. On the one hand, process monitoring is required for the early detection of batches with an out-of-spec end-point product quality, with the aim of minimising the amount of out-of-spec batches produced. On the other hand, process monitoring is carried out to detect anomalies in the process operating conditions with the aim of troubleshooting the process, even if the end-point product is within specification.

Specifically, the following research areas were investigated:

1. the assessment of the benefits of the application of batch process monitoring techniques on a real and complex (multi-unit) industrial process;
2. the development of automated approaches for batch alignment for end-point quality estimation;
3. the development of guidelines for the implementation of batch alignment-free methodologies.

Table C.1 summarizes the main achievements of this Dissertation, with indication of example applications, type of data used, and references where the results have been discussed.
In Chapter 4, the application of data analytics techniques for troubleshooting a complex industrial batch process highlighted how batch alignment is an important and time consuming preprocessing step affecting the performance of the analysis.
This preliminary work allowed to carry out an **assessment of the advantages of the application of batch process monitoring techniques** on a real and complex industrial process, resulting in industrially-relevant results in terms of efficiency and productivity improvements.
In more detail, in Chapter 4 data analytics techniques were employed on a semi-batch manufacturing process and coupled with engineering understanding for process troubleshooting. The batch-wise unfolding multiway principal component analysis model developed allowed to uncover the existence of an abnormal behaviour affecting 40% of the batches in historical manufacturing campaigns. This behaviour went almost unnoticed as batches did not terminate unsuccessfully, but simply took longer to complete than the others. However, this anomaly increased the utility and energy expenditure per unit of product manufactured. Furthermore, being the process under investigation a bottleneck of the overall process, the longer average batch length reduced the productivity of the overall process. Data analytics was central to identify the cause of the abnormal behaviour in the anomalous intervention of one interlock in the reactor safety system, which caused vacuum in the reactor to be broken by nitrogen blanketing under particular conditions, requiring a subsequent vacuum reinstatement to recover the process conditions and to end the batch successfully. Reconfiguration of the safety interlock system allowed to shorten the average batch length by 29%, and the overall process cycle time by 8%. Furthermore, an 11% reduction on the nitrogen consumption was obtained. The development of the data analysis model for this case study highlighted that batch matching is an important and complex pre-processing step that affects the performance of the analysis. This finding was instrumental in the development of techniques that allow the practitioner to perform this step easily and to avoid it altogether when not absolutely necessary. In fact, artificially reducing all batches to a common length (i.e. batch alignment) is usually done by trial and error, is time consuming and requires prior process knowledge.

With respect to cases in which it is necessary to apply batch alignment for process monitoring, in Chapter 5 an **automated methodology for batch alignment in batch end-point product quality estimation** is proposed.
The proposed methodology was developed to assist the practitioner at maximizing the end-point quality monitoring model performance. The methodology retains the efficacy of a straightforward trajectory synchronization method, such as the traditional indicator variable method. It improves upon this method in two ways: *i*) by automatically performing the

partitioning into phases, and *ii*) by selecting the most appropriate indicator variable for each phase. This is done simultaneously rather than separately, utilizing an optimization framework to maximize the performance of a model for product quality assessment that is to be constructed using the available datasets. Differently from classic indicator variable and from advanced synchronization methodologies such as dynamic time warping, correlation optimized warping and multisynchro, the proposed methodology is process agnostic as it does not require identifying either a reference batch or a reference variable. The proposed data preprocessing methodology was tested on two datasets, one from an industrial process and one from a simulated one. The performance of the resulting product quality assessment model when the available data were preprocessed with the proposed methodology and with other synchronization strategies were compared. The proposed methodology always led to the best performance, both when the model was used as a soft sensor to estimate the product end-point quality and when the model was used as a classifier to discriminate between on-spec and off-spec products. From the computational side, the proposed methodology performance outperforms other advanced methodologies when they become computationally intensive, i.e. when the number of samples per batch increases.

Finally, when process monitoring is carried out with the aim of detecting anomalies in the process operating conditions, one can use a methodology that does not require batch alignment, namely, the assumption-free monitoring methodology proposed a few years ago in the literature. However, effective implementation of this methodology is challenging due to the lack of sufficient documentation and of a clear fault detection and diagnosis procedure. In Chapter 6, a **set of guidelines for the implementation of this batch alignment-free monitoring methodology** is proposed to allow the practitioner to implement the methodology in a straightforward manner. Light was shed on the optimization problem underlying the calibration of the assumption-free model and an algorithmic implementation to solve the problem in a particular case of industrial interest was provided. Furthermore, clear indications for performing fault detection were provided together with a novel methodology for fault diagnosis. The assumption-free monitoring methodology implemented according to the proposed guidelines was tested on five datasets, 2 with even-length and 3 with uneven-length batches, coming from different industrial sectors and processes of different nature (chemical, biological and physical). The performance of the methodology was compared with a traditional methodology for batch process monitoring, based on batch-wise unfolding of the data matrix. The proposed methodology consistently outperformed the traditional one, especially in cases when the number of time samples of the batches is large, as it is not affected by a deterioration of the SPE statistic, typically observed when applying batch-wise unfolding to datasets with a large number of samples per batch.

Table C.1: *Summary of the main achievements of this Dissertation, with indication of their relevant applications, the origin of the data used and related references*

Chapter	Main achievement	Application	Data origin	Reference
Chapter 4	Assessment of the effectiveness of multivariate batch process monitoring in an industrial setting	• Specialty chemical manufacturing	Industrial	Sartori F., Zuecco F., Facco, P., Bezzo F., Barolo M. (2022), Data Analytics Can Help Reduce Energy Consumption in the Industrial Manufacturing of Specialty Chemicals. *Chem. Eng. Trans.* **96**, 229-234 Sartori F., Zuecco F., Facco P., Bezzo F., Barolo M. (2022), Coupling machine learning and engineering judgment to reduce the cycle time of an industrial batch process, Oral presentation at: *GRICU 2022: Centralità dell'Ingegneria Chimica in un Mondo che cambia, 03-06 July, 2022*, (Ischia, NA, Italy) Sartori F., Zuecco F., Facco P., Bezzo F., Barolo M. (2022), Data Analytics Can Help Reduce Energy Consumption in the Industrial Manufacturing of Specialty Chemicals. Oral presentation at: *1st International Conference on Energy, Environment and Digital Transition (E2DT), October 23-26 2022* (Milano, Italy)
Chapter 5	Development of an automated methodology for batch alignment for batch end-point estimation	• Specialty chemical manufacturing • Penicillin fermentation	Industrial and simulated	Sartori F., Facco, P., Zuecco F., Bezzo F., Barolo M. (2023), Optimal indicator-variable approach for trajectory synchronization in uneven-length multiphase batch processes. *Ind. Eng. Chem. Res.* **62**, 18511-18525. Sartori F., Zuecco F., Facco P., Bezzo F., Barolo M. (2023), An automated pre-processing framework for uneven-length multiphase batch processes. Oral presentation at: *11th Colloquium Chemometricum Mediterraneum, June 27-30 2023* (Padova, Italy)
Chapter 6	Development of guidelines for the application of alignment-free methodologies for batch process monitoring	• SBR polymerization • LDPE polymerization • Baker's yeast fermentation • Penicillin fermentation • Herbicide crystals drying	Industrial and simulated	Sartori F., Facco, P., Bezzo F., Barolo M. (2023), On the application of assumption-free modelling for multivariate statistical batch process monitoring. *In preparation.*

Furthermore, the adaptive nature of the control limits in the assumption-free methodology allowed to outperform the traditional method even in situations where a large batch-to-batch variability is observed. The batch-wise unfolding based methodology outperforms the assumption-free methodology in cases when the dynamic of the process is compressed in a very small fraction of the variables time trajectories. The assumption-free methodology is also less computationally expensive than the traditional methodology by definition, as it does not require batch alignment. Furthermore, the proposed fault diagnosis methodology consistently outperformed the methodology actually used for fault diagnosis in the assumption-free technique.

Some **areas of further investigation** can be discussed at this point.

- The methodology proposed in Chapter 5 has been applied to two processes, an industrial and a simulated one, however, it could be relevant extending its application to different industrial sectors and test it on a variety of case studies, in particular, the methodology proposed in Chapter 5 could be extended to nonlinear models for estimating the final product quality in order to handle strong process nonlinearities, where the proposed linear approach is deemed to fail in most practical situations;
- different surrogates can be used other than the radial basis function one for surrogate optimization in the methodology proposed in Chapter 5 and a comparison between the performance of different surrogates can be carried out to further improve the methodology performance;
- the guidelines proposed for the implementation of the methodology proposed in Chapter 6 propose an explicit optimization problem to be solved for the development of an assumption-free model, however, the algorithm proposed for its solution solves it in a particular case of industrial interest. The development of an algorithm based on appropriate heuristics for the calibration of the assumption-free model for any number of principal components can enhance the model performance; Furthermore, the assumption-free model is actually based on the construction of a grid with fixed resolution for analyzing the multivariate trajectory of the process under study. A method for exploring the latent space in an adaptive manner both in terms of geometry and size of the segments in which it is subdivided can improve the performance of the technique, especially in cases where the dynamics of the process under study is condensed in a relatively small portion of the trajectory.

References

Achterberg, T., Koch, T., Martin, A., 2005. Branching rules revisited. Oper. Res. Lett. 33, 42–54. https://doi.org/10.1016/j.orl.2004.04.002

Allen, N.S., Edge, M., 2021a. Perspectives on additives for polymers. 1. Aspects of stabilization. J. Vinyl Addit. Technol. 27, 5–27. https://doi.org/10.1002/vnl.21807

Allen, N.S., Edge, M., 2021b. Perspectives on additives for polymers. Part 2. Aspects of photostabilization and role of fillers and pigments. J. Vinyl Addit. Technol. 27, 211–239. https://doi.org/10.1002/vnl.21810

American Chemistry Council, 2022. 2022 Guide To the Business of Chemistry.

Amigo, J.M., Skov, T., Bro, R., Coello, J., Maspoch, S., 2008. Solving GC-MS problems with PARAFAC2. TrAC - Trends Anal. Chem. 27, 714–725. https://doi.org/10.1016/j.trac.2008.05.011

Ballabio, D., Consonni, V., 2013. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Anal. Methods 5, 3790–3798. https://doi.org/10.1039/c3ay40582f

Bano, G., 2023. Successful implementation of quality by design along the product life cycle – industrial use cases [WWW Document]. EFCE Spotlight Talks. URL https://www.youtube.com/watch?v=TbWF6vw0_DA (accessed 6.23.23).

Bano, G., Wang, Z., Facco, P., Bezzo, F., Barolo, M., Ierapetritou, M., 2018. A novel and systematic approach to identify the design space of pharmaceutical processes. Comput. Chem. Eng. 115, 309–322. https://doi.org/10.1016/j.compchemeng.2018.04.021

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. J. Chemom. 17, 166–173. https://doi.org/10.1002/cem.785

Barton, M., Duran-Villalobos, C.A., Lennox, B., 2021. Multivariate batch to batch optimisation of fermentation processes to improve productivity. J. Process Control 108, 148–156. https://doi.org/10.1016/j.jprocont.2021.11.007

BASF, 2022. Outlook for the Chemical Industry [WWW Document]. BASF Rep. 2022. URL https://report.basf.com/2022/en/managements-report/forecast/economic-environment/chemical-industry.html (accessed 8.2.23).

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. Comput. Chem. Eng. 108, 250–267. https://doi.org/10.1016/j.compchemeng.2017.09.017

Biancolini, M.E., 2017. Fast Radial Basis Functions for Engineering Applications. Springer International Publishing, Heidelberg.

Bieler, P.S., Fischer, U., Hungerbühler, K., 2004. Modeling the energy consumption of chemical batch plants: Bottom-up approach. Ind. Eng. Chem. Res. 43, 7785–7795. https://doi.org/10.1021/ie049641j

Birol, G., Ündey, C., Çinar, A., 2002. A modular simulation package for fed-batch fermentation: Penicillin production. Comput. Chem. Eng. 26, 1553–1565. https://doi.org/10.1016/S0098-1354(02)00127-8

Bonvin, D., Srinivasan, B., Hunkeler, D., 2006. Control and optimization of batch processes: Improvement of process operation in the production of specialty chemicals. IEEE Control Syst. 26, 34–45. https://doi.org/10.1109/MCS.2006.252831

Bro, R., Kjeldahl, K., Smilde, A.K., Kiers, H.A.L., 2008. Cross-validation of component models: A critical look at current methods. Anal. Bioanal. Chem. 390, 1241–1251. https://doi.org/10.1007/s00216-007-1790-1

Bro, R., Smilde, A.K., 2014. Principal component analysis. Anal. Methods 6, 2812–2831. https://doi.org/10.1039/c3ay41907j

Broadhead, T.O., Hamielec, A.E., MacGregor, J.F., 1985. Dynamic modelling of the batch, semi-batch and continuous production of styrene/butadiene copolymers by emulsion polymerization. Die Makromol. Chemie 10, 105–128.

Brown, J.D., 2009. Choosing the Right Number of Components of Factors in PCA and EFA. JALT Test. Eval. SIG Newsl. 13, 19–23.

Brunner, V., Klöckner, L., Kerpes, R., Geier, D.U., Becker, T., 2020. Online sensor validation in sensor networks for bioprocess monitoring using swarm intelligence. Anal. Bioanal. Chem. 412, 2165–2175. https://doi.org/10.1007/s00216-019-01927-7

BusinessWire, 2023. Global Specialty Chemicals Market [WWW Document]. URL https://www.businesswire.com/news/home/20230518005615/en/Global-Specialty-Chemicals-Market-Report-2023-A-738.23-Billion-Market-in-2022---Forecasts-to-2028---Rising-Demand-For-Sustainable-Specialty-Chemicals-High-Performance-Materials---ResearchAndMarkets. (accessed 7.11.23).

Camacho, J., Ferrer, A., 2014. Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: Practical aspects. Chemom. Intell. Lab. Syst. 131, 37–50. https://doi.org/10.1016/j.chemolab.2013.12.003

Camacho, J., Picó, J., Ferrer, A., 2008a. Multi-phase analysis framework for handling batch process data. J. Chemom. 22, 632–643. https://doi.org/10.1002/cem.1151

Camacho, J., Picó, J., Ferrer, A., 2008b. Bilinear modelling of batch processes. Part I: Theoretical discussion. J. Chemom. 22, 299–308. https://doi.org/10.1002/cem.1113

Carlsson, D.J., Jensen, J.P.T., Wiles, D.M., 1984. Antioxidant mechanisms of hindered amine light stabilizers. Die Makromol. Chemie 8, 79–88. https://doi.org/10.1002/macp.1984.020081984107

Cefic, 2023. Cefic Facts & Figures 2023.

Chen, G., Zhang, K., Xue, X., Zhang, L., Yao, C., Wang, J., Yao, J., 2022. A radial basis function surrogate model assisted evolutionary algorithm for high-dimensional expensive optimization problems. Appl. Soft Comput. 116, 108353. https://doi.org/10.1016/j.asoc.2021.108353

Chiang, L.H., Braun, B., Wang, Z., Castillo, I., 2022. Towards artificial intelligence at scale in the chemical industry. AIChE J. https://doi.org/10.1002/aic.17644

Chiang, L.H., Leardi, R., Pell, R.J., Seasholtz, M.B., 2006. Industrial experiences with multivariate statistical analysis of batch process data. Chemom. Intell. Lab. Syst. 81, 109–119. https://doi.org/10.1016/j.chemolab.2005.10.006

Chiang, L.H., Russell, E.L., Braatz, R.D., 2001. Fault Detection and Diagnosis in Industrial Systems. Springer-Verlag, London.

Cui, Z.H., Xia, G., Chen, W.G., Jiang, H., Yang, L., Zuo, Z.W., 2020. Synthesis of novel multifunctional photostabilizers containing UVA and HALS moieties and their effects on polymers and dyes. J. Vinyl Addit. Technol. 26, 259–267. https://doi.org/10.1002/vnl.21740

Dunn, K., 2019. Geometric Explanation of PCA [WWW Document]. Process Improv. Using Data. URL https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/geometric-explanation-of-pca (accessed 6.20.19).

Eriksson, L., 2021. Batch Process Modeling – Step-By-Step Guide. Umeå.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., Wold, S., 2006. Multi- and megavariate data analysis. Umetrics, Umeå.

Ferrer, A., 2020. Discussion of "A review of data science in business and industry and a future view," by Grazia Vicario and Shirley Coleman. Appl. Stoch. Model. Bus. Ind. 1–7.

https://doi.org/10.1002/asmb.2506

Fransson, M., Folestad, S., 2006. Real-time alignment of batch process data using COW for on-line process monitoring. Chemom. Intell. Lab. Syst. 84, 56–61. https://doi.org/10.1016/j.chemolab.2006.04.020

Furstenau, L.B., Sott, M.K., Kipper, L.M., MacHado, E.L., Lopez-Robles, J.R., Dohan, M.S., Cobo, M.J., Zahid, A., Abbasi, Q.H., Imran, M.A., 2020. Link between Sustainability and Industry 4.0: Trends, Challenges and New Perspectives. IEEE Access 8, 140079–140096. https://doi.org/10.1109/ACCESS.2020.3012812

García-Muñoz, S., Kourti, T., MacGregor, J.F., Mateos, A.G., Murphy, G., 2003. Troubleshooting of an industrial batch process using multivariate methods. Ind. Eng. Chem. Res. 42, 3592–3601. https://doi.org/10.1021/ie0300023

García-Muñoz, S., Polizzi, M., Prpich, A., Strain, C., Lalonde, A., Negron, V., 2011. Experiences in batch trajectory alignment for pharmaceutical process improvement through multivariate latent variable modelling. J. Process Control 21, 1370–1377. https://doi.org/10.1016/j.jprocont.2011.07.013

Geladi, P., Kowalski, B.R., 1986. Partial Least-Squares regression: a tutorial. Anal. Chim. Acta 1–17.

Gilbert, R.O., 1987. Statistical Methods for Environment Pollution Monitoring, 1st ed. John Wiley and Sons Inc., New York.

Golshan, M., MacGregor, J.F., Bruwer, M.J., Mhaskar, P., 2010. Latent Variable Model Predictive Control (LV-MPC) for trajectory tracking in batch processes. J. Process Control 20, 538–550. https://doi.org/10.1016/j.jprocont.2010.01.007

González-Martínez, J.M., Camacho, J., Ferrer, A., 2018. MVBatch: A matlab toolbox for batch process modeling and monitoring. Chemom. Intell. Lab. Syst. 183, 122–133. https://doi.org/10.1016/j.chemolab.2018.11.001

González-Martínez, José M., de Noord, O.E., Ferrer, A., 2014. Multisynchro: A novel approach for batch synchronization in scenarios of multiple asynchronisms. J. Chemom. 28, 462–475. https://doi.org/10.1002/cem.2620

González-Martínez, J.M., Ferrer, A., Westerhuis, J.A., 2011. Real-time synchronization of batch trajectories for on-line multivariate statistical process control using Dynamic Time Warping. Chemom. Intell. Lab. Syst. 105, 195–206. https://doi.org/10.1016/j.chemolab.2011.01.003

González-Martínez, J. M., Vitale, R., De Noord, O.E., Ferrer, A., 2014. Effect of synchronization on bilinear batch process modeling. Ind. Eng. Chem. Res. 53, 4339–4351. https://doi.org/10.1021/ie402052v

Guillet, J.E., 1972. Fundamental Processes in the Uv Degradation and Stabilization of Polymers, Chemical Transformations of Polymers. International Union of Pure and Applied Chemistry. https://doi.org/10.1016/b978-0-408-70310-9.50012-5

Guisinger, A., Ghorashi, B., 2004. Agile manufacturing practices in the specialty chemical industry An overview of the trends and results of a specific case study. Int. J. Oper. Prod. Manag. 24, 625–635. https://doi.org/10.1108/01443570410538140

Guo, R., Jin, Y., 2019. Phase Identification and Online Monitoring for the Uneven Batch Processes. IEEE Access 7, 81351–81363. https://doi.org/10.1109/ACCESS.2019.2919167

Gutmann, H.M., 2001. A Radial Basis Function Method for Global Optimization. J. Glob. Optim. 19, 201–227. https://doi.org/10.1023/A:1011255519438

He, Q.P., Wang, J., 2011. Statistics Pattern Analysis: A New Process Monitoring Framework and its Application to Semiconductor Batch Processes. AIChE J. 57, 107–121. https://doi.org/10.1002/aic

Höskuldsson, A., 1996. Experimental design and priority PLS regression. J. Chemom. 10, 637–

668.

Höskuldsson, A., 1988. PLS Regression Methods. J. Chemom. 2, 211–228.

Indahl, U.G., 2014. Towards a complete identification of orthogonal variation in multiple regression froma PLS1 modeling point of view: including OPLS by a change of orthogonal basis. J. Chemom. 28, 508–517.

Iske, A., 2002. Scattered Data Modelling Using Radial Basis Functions, in: Tutorials on Multiresolution in Geometric Modelling. Springer-Verlag, Berlin, Heidelberg, pp. 205–242. https://doi.org/https://doi.org/10.1007/978-3-662-04388-2_9

Jackson, J.E., 1991. A User's Guide to Principal Components. https://doi.org/10.2307/2583020

Ji, C., Sun, W., 2022. A Review on Data-Driven Process Monitoring Methods: Characterization and Mining of Industrial Data. Processes 10. https://doi.org/10.3390/pr10020335

Joe Qin, S., 2003. Statistical process monitoring: basics and beyond. J. Chemom. 17, 480–502. https://doi.org/10.1002/cem.800

Jollife, I.T., Cadima, J., 2016. Principal component analysis: A review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374. https://doi.org/10.1098/rsta.2015.0202

Kagermann, H., Wahlster, W., 2022. Ten Years of Industrie 4.0. Sci 4. https://doi.org/10.3390/sci4030026

Kassidas, A., MacGregor, J.F., Taylor, P.A., 1998. Synchronization of Batch Trajectories Using Dynamic Time Warping. AIChE J. 44, 864–875. https://doi.org/10.1002/aic.690440412

Kourti, T., 2003. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. J. Chemom. 17, 93–109. https://doi.org/10.1002/cem.778

Kourti, T., Lee, J., Macgregor, J.F., 1996. Experiences with industrial applications of projection methods for multivariate statistical process control. Comput. Chem. Eng. 20, 745–750. https://doi.org/10.1016/0098-1354(96)00132-9

Krause, D., Hussein, M.A., Becker, T., 2015. Online monitoring of bioprocesses via multivariate sensor prediction within swarm intelligence decision making. Chemom. Intell. Lab. Syst. 145, 48–59. https://doi.org/10.1016/j.chemolab.2015.04.012

Lakshminarayanan, S., Gudi, R.D., Shah, S.L., Nandakumar, K., 1996. Monitoring Batch Processes Using Multivariate Statistical Tools: Extensions and Practical Issues. IFAC Proc. Vol. 29, 6037–6042. https://doi.org/10.1016/s1474-6670(17)58648-6

Lasi, H., Fettke, P., Kemper, H.G., Feld, T., Hoffmann, M., 2014. Industry 4.0. Bus. Inf. Syst. Eng. 6, 239–242. https://doi.org/10.1007/s12599-014-0334-4

Lei, F., Rotboll, M., Jorgensen, S.B., 2001. A biochemically structured model for Saccharomyces cerevisiae. J. Biotechnol. 88, 205–221. https://doi.org/10.1016/S0168-1656(01)00269-3

Louwerse, D.J., Smilde, A.K., 2000. Multivariate statistical process control of batch processes based on three-way models. Chem. Eng. Sci. 55, 1225–1235. https://doi.org/10.1016/S0009-2509(99)00408-X

Lu, B., Xu, S., Stuber, J., Edgar, T.F., 2016. Constrained selective dynamic time warping of trajectories in three dimensional batch data. Chemom. Intell. Lab. Syst. 159, 138–150. https://doi.org/10.1016/j.chemolab.2016.10.005

Lu, N., Gao, F., Yang, Y., Wang, F., 2004. PCA-based modeling and on-line monitoring strategy for uneven-length batch processes. Ind. Eng. Chem. Res. 43, 3343–3352. https://doi.org/10.1021/ie030736f

Lu, Y., 2017. Industry 4.0: A survey on technologies, applications and open research issues. J. Ind. Inf. Integr. 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Luo, L., Bao, S., Gao, Z., 2015. Quality prediction based on HOPLS-CP for batch processes. Chemom. Intell. Lab. Syst. 143, 28–39. https://doi.org/10.1016/j.chemolab.2015.02.010

Luo, L., Bao, S., Mao, J., Tang, D., 2016. Phase Partition and Phase-Based Process Monitoring Methods for Multiphase Batch Processes with Uneven Durations. Ind. Eng. Chem. Res. 55, 2035–2048. https://doi.org/10.1021/acs.iecr.5b03993

MacGregor, J.F., Cinar, A., 2012. Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. Comput. Chem. Eng. 47, 111–120.

MacGregor, J.F., Jaeckle, C., Kiparissides, C., Koutoudi, M., 1994. Process monitoring and diagnosis by multiblock PLS methods. AIChE J. 40, 826–838. https://doi.org/10.1002/aic.690400509

MacGregor, J.F., Kourti, T., 1995. STATISTICAL PROCESS CONTROL OF MULTIVARIATE PROCESSES. Control Eng. Pract. 3, 403–414.

Magán-Carrión, R., Pulido-Pulido, F., Camacho, J., García-Teodoro, P., 2013. Tampered data recovery in WSNs through dynamic PCA and variable routing strategies. J. Commun. 8, 738–750. https://doi.org/10.12720/jcm.8.11.738-750

McKinsey, 2020. The impact of COVID-19 on the global petrochemical industry [WWW Document]. URL https://www.mckinsey.com/industries/chemicals/our-insights/the-impact-of-covid-19-on-the-global-petrochemical-industry (accessed 8.2.23).

Miller, P., Swanson, R.E., Heckler, C.E., 1998. Contribution plots: a missing link in multivariate quality control. J. Appl. Math. Comput. Sci.

Moulijn, J.A., Makkee, M., Van Diepen, A.E., 2014. Chemical Process Technology, 2nd ed. John Wiley and Sons Inc., Chichester.

Mousavi-Fakhrabadi, S.H., Ahmadi, S., Arabi, H., 2022. Mixing of hindered amine-grafted polyolefin elastomers with LDPE to enhance its long-term weathering and photo-stability. Polym. Degrad. Stab. 198, 109882. https://doi.org/10.1016/j.polymdegradstab.2022.109882

Mueen, A., Keogh, E., 2016. Extracting optimal performance from dynamic time warping. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-Augu, 2129–2130. https://doi.org/10.1145/2939672.2945383

Neogi, D., Schlags, C.E., 1998. Multivariate statistical analysis of an emulsion batch process. Ind. Eng. Chem. Res. 37, 3971–3979. https://doi.org/10.1021/ie980243o

Nielsen, N.P.V., Carstensen, J.M., Smedsgaard, J., 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J. Chromatogr. A 805, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Nomikos, P., 1996. Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis World Batch Forum, Toronto, May 1996. ISA Trans. 35, 259–266. https://doi.org/10.1016/s0019-0578(96)00035-3

Nomikos, P., MacGregor, J.F., 1995a. Multivariate SPC charts for monitoring batch processes. Technometrics 37, 41–59. https://doi.org/10.1080/00401706.1995.10485888

Nomikos, P., MacGregor, J.F., 1995b. Multi-way partial least squares in monitoring batch processes. Chemom. Intell. Lab. Syst. 30, 97–108. https://doi.org/10.1016/0169-7439(95)00043-7

Nomikos, P., MacGregor, J.F., 1994. Monitoring Batch Processes Using Multiway Principal Component Analysis. AIChE J. 40, 1361–1375.

OECD, 2022. The supply of critical raw materials endangered by Russia's war on Ukraine [WWW Document]. URL https://www.oecd.org/ukraine-hub/policy-responses/the-supply-of-critical-raw-materials-endangered-by-russia-s-war-on-ukraine-e01ac7be/

Oláh, J., Aburumman, N., Popp, J., Khan, M.A., Haddad, H., Kitukutha, N., 2020. Impact of

industry 4.0 on environmental sustainability. Sustain. 12, 1–21. https://doi.org/10.3390/su12114674

Pilario, K.E., Shafiee, M., Cao, Y., Lao, L., Yang, S.H., 2020. A review of kernel methods for feature extraction in nonlinear process monitoring. Processes 8, 1–47. https://doi.org/10.3390/pr8010024

Qin, S.J., 2012. Survey on data-driven industrial process monitoring and diagnosis. Annu. Rev. Control 36, 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004

Queipo, N. V., Haftka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Kevin Tucker, P., 2005. Surrogate-based analysis and optimization. Prog. Aerosp. Sci. 41, 1–28. https://doi.org/10.1016/j.paerosci.2005.02.001

Rabek, J.F., 1990. Photostabilization of polymers: Principles and Applications, 1st ed. Elsevier Science Publishers LTD, London, New York.

Raich, A., Çinar, A., 1996. Statistical Process Monitoring and Disturbance Diagnosis in Multivariable Continuous Processes. AIChE J. 42, 995–1009. https://doi.org/10.1002/aic.690420412

Rato, T.J., Blue, J., Pinaton, J., Reis, M.S., 2017. Translation-Invariant Multiscale Energy-Based PCA for Monitoring Batch Processes in Semiconductor Manufacturing. IEEE Trans. Autom. Sci. Eng. 14, 894–904. https://doi.org/10.1109/TASE.2016.2545744

Rato, T.J., Rendall, R., Gomes, V., Chin, S.T., Chiang, L.H., Saraiva, P.M., Reis, M.S., 2016. A Systematic Methodology for Comparing Batch Process Monitoring Methods: Part I-Assessing Detection Strength. Ind. Eng. Chem. Res. 55, 5342–5358. https://doi.org/10.1021/acs.iecr.5b04851

Rato, T.J., Rendall, R., Gomes, V., Saraiva, P.M., Reis, M.S., 2018. A Systematic Methodology for Comparing Batch Process Monitoring Methods: Part II - Assessing Detection Speed. Ind. Eng. Chem. Res. 57, 5338–5350. https://doi.org/10.1021/acs.iecr.7b04911

Reis, M.S., Gins, G., 2017. Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. Processes 5. https://doi.org/10.3390/pr5030035

Reis, M.S., Kenett, R., 2018. Assessing the value of information of data-centric activities in the chemical processing industry 4.0. AIChE J. 64, 3868–3881. https://doi.org/10.1002/aic.16203

Reis, M.S., Rendall, R., Rato, T.J., Martins, C., Delgado, P., 2021. Improving the sensitivity of statistical process monitoring of manifolds embedded in high-dimensional spaces: The truncated-Q statistic. Chemom. Intell. Lab. Syst. 215, 104369. https://doi.org/10.1016/j.chemolab.2021.104369

Rendall, R., Chiang, L.H., Reis, M.S., 2019. Data-driven methods for batch data analysis – A critical overview and mapping on the complexity scale. Comput. Chem. Eng. 124, 1–13. https://doi.org/10.1016/j.compchemeng.2019.01.014

Rendall, R., Lu, B., Castillo, I., Chin, S.T., Chiang, L.H., Reis, M.S., 2017a. A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes. Ind. Eng. Chem. Res. 56, 8590–8605. https://doi.org/10.1021/acs.iecr.6b04553

Rendall, R., Lu, B., Castillo, I., Chin, S.T., Chiang, L.H., Reis, M.S., 2017b. Profile-driven Features for Offline Quality Prediction in Batch Processes. Comput. Aided Chem. Eng. 40, 1501–1506.

Rosipal, R., Krämer, N., 2006. Overview and Recent Advances in Partial Least Squares., in: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (Eds.), Subspace, Latent Structure and Feature Selection. SLSFS 2005. Springer-Verlag, Berlin, Heidelberg.

Rothwell, S.G., Martin, E.B., Morris, A.J., 1998. Comparison of Methods for Dealing with Uneven Length Batches. IFAC Proc. Vol. 31, 387–392. https://doi.org/10.1016/s1474-6670(17)40216-3

Sanchez-Fernández, A., Fuente, M.J., Sainz-Palmero, G.I., 2015. Fault detection in wastewater

treatment plants using distributed PCA methods. IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA 2015-Octob, 1–7. https://doi.org/10.1109/ETFA.2015.7301504

Schaller, C., Rogez, D., Braig, A., 2009. Hindered amine light stabilizers in pigmented coatings. J. Coatings Technol. Res. 6, 81–88. https://doi.org/10.1007/s11998-008-9130-8

Shang, C., You, F., 2019. Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era. Engineering 5, 1010–1016. https://doi.org/10.1016/j.eng.2019.01.019

She, C., Wang, Z., Sun, F., Liu, P., Zhang, L., 2020. Battery Aging Assessment for Real-World Electric Buses Based on Incremental Capacity Analysis and Radial Basis Function Neural Network. IEEE Trans. Ind. Informatics 16, 3345–3354. https://doi.org/10.1109/TII.2019.2951843

Sirimanne, S.N., 2022. What is "Industry 4.0" and what will it mean for developing countries? [WWW Document]. UNCTAD. URL https://unctad.org/news/blog-what-industry-40-and-what-will-it-mean-developing-countries (accessed 7.20.23).

Smit, J., Kreutzer, S., Moeller, C., Carlberg, M., 2016. Industry 4.0: study for the ITRE Committee. Brussels.

SOCMA, 2023. Specialty Chemistry [WWW Document]. URL https://www.socma.org/about/specialty-chemistry/

Sun, W., Meng, Y., Palazoglu, A., Zhao, J., Zhang, H., Zhang, J., 2011. A method for multiphase batch process monitoring based on auto phase identification. J. Process Control 21, 627–638. https://doi.org/10.1016/j.jprocont.2010.12.003

The Mathworks, 2020. MATLAB version 9.8 (R2020a).

Tian, K., Wu, L., Min, S., Bro, R., 2018. Geometric search: A new approach for fitting PARAFAC2 models on GC-MS data. Talanta 185, 378–386. https://doi.org/10.1016/j.talanta.2018.03.088

Tomasi, G., Van Den Berg, F., Andersson, C., 2004. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. J. Chemom. 18, 231–241. https://doi.org/10.1002/cem.859

Ündey, C., Ertunç, S., Çinar, A., 2003. Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. Ind. Eng. Chem. Res. 42, 4645–4658. https://doi.org/10.1021/ie0208218

Vanlaer, J., Van der Kerkhof, P., Gins, G., Van Impe, J.F.M., 2012. The Influence of Input and Output Measurement Noise on Batch-End Quality Prediction with Partial Least Squares, in: Advances in Data Mining. Applications and Theoretical Aspects: 12th Industrial Conference, ICDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 12. Springer Berlin Heidelberg, pp. 121–135.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., Yin, K., 2003a. A review of process fault detection and diagnosis Part I: Quantitative model-based methods. Comput. Chem. Eng. 27, 327–346. https://doi.org/10.1016/s0098-1354(02)00162-x

Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N., 2003b. A review of fault detection and diagnosis. Part III: Process history based methods. Comput. Chem. Eng. 27, 327–346.

Wan, J., Marjanovic, O., Lennox, B., 2014. Uneven batch data alignment with application to the control of batch end-product quality. ISA Trans. 53, 584–590. https://doi.org/10.1016/j.isatra.2013.12.020

Westad, F., 2020. Assumption-free modeling and monitoring of batch processes [WWW Document]. YouTube. URL https://www.youtube.com/watch?v=BbkX0aq3rpM (accessed 6.13.23).

Westad, F., Gidskehaug, L., Swarbrick, B., Flåten, G.R., 2015. Assumption free modeling and monitoring of batch processes. Chemom. Intell. Lab. Syst. 149, 66–72.

https://doi.org/10.1016/j.chemolab.2015.08.022

Wise, B.M., Gallagher, N.B., 1996. The process chemometrics approach to process monitoring and fault detection. J. Process Control 6, 329–348.

Wold, S., 1978. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics 20, 397–405. https://doi.org/10.1080/00401706.1978.10489693

Wold, S., Kettaneh, N., Fridén, H., Holmberg, A., 1998. Modelling and diagnostics of batch processes and analogous kinetic experiments. Chemom. Intell. Lab. Syst. 44, 331–340. https://doi.org/10.1016/S0169-7439(98)00162-2

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method, in: Matrix Pencils. pp. 286–293.

Wold, S., Ruhe, A., Wold, H., Dunn, W., 1984. The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5, 735–743.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130.

Xu, Y., Zomer, S., Brereton, R.G., 2006. Support vector machines: A recent method for classification in chemometrics. Crit. Rev. Anal. Chem. 36, 177–188. https://doi.org/10.1080/10408340600969486

Yu, H., Augustijn, D., Bro, R., 2021. Accelerating PARAFAC2 algorithms for non-negative complex tensor decomposition. Chemom. Intell. Lab. Syst. 214, 104312. https://doi.org/10.1016/j.chemolab.2021.104312

Zhang, S., Zhao, C., Gao, F., 2018. Two-directional concurrent strategy of mode identification and sequential phase division for multimode and multiphase batch process monitoring with uneven lengths. Chem. Eng. Sci. 178, 104–117. https://doi.org/10.1016/j.ces.2017.12.025

Zhao, C., 2014. A quality-relevant sequential phase partition approach for regression modeling and quality prediction analysis in manufacturing processes. IEEE Trans. Autom. Sci. Eng. 11, 983–991. https://doi.org/10.1109/TASE.2013.2287347

Zheng, X., Wang, M., Ordieres-Meré, J., 2018. Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0. Sensors (Switzerland) 18. https://doi.org/10.3390/s18072146

Zhong, D.Y., Wang, L.G., Bi, L., 2019. Implicit surface reconstruction based on generalized radial basis functions interpolant with distinct constraints. Appl. Math. Model. 71, 408–420. https://doi.org/10.1016/j.apm.2019.02.026

Zhou, M., Wong, M.H., 2008. Efficient online subsequence searching in data streams under dynamic time warping distance. Proc. - Int. Conf. Data Eng. 00, 686–695. https://doi.org/10.1109/ICDE.2008.4497477

Acknowledgements

I would like to express my heartfelt gratitude to all those who have contributed to the completion of this Dissertation. The journey from the inception of this research to its culmination has been a long and challenging one, and I could not have reached this point without the support and encouragement of many individuals and institutions.

First and foremost, I am deeply thankful to my advisor, Prof. Massimiliano Barolo, whose unwavering guidance, patience, and expertise have been instrumental in shaping the direction of this research. Your mentorship has been invaluable, and I am grateful for the time you have dedicated to helping me refine my ideas and navigate the complexities of academic research.
I would also like to thank Prof. Fabrizio Bezzo and Prof. Pierantonio Facco, for their advice and mentoring during my PhD. Thank you for being a guidance in my scientific and personal choices. Sincere thanks go to past and present members of CAPE-Lab: Christopher, Luca, Francesco, Alberto, Elia, Francesca, Andrea, Daniel, Beatriz, Margherita for having provided a supportive, friendly and intellectually stimulating environment.
Warm thanks go to Gianmarco, for being a good colleague and a special friend that shared with me the best part of these years.
Grateful appreciation go to Federico for the collaboration we established, especially during my stay at the BASF Italia Pontecchio Marconi site. I am also deeply grateful to Stefan for mentoring me during my stay as a Visiting PhD Student at BASF Lampertheim and for giving a huge contribution to my professional growth in such a limited time. Wholehearted thanks go also to Peter, Tim and all the OpEx team in BASF Lampertheim for your warm welcome in Lampertheim.
Special thanks go to Federica, your presence in my life has been a source of joy, support and inspiration and have enriched my life in countless ways, I am fortunate to have you by my side.
Finally I would like to thank my family, your support, love and sacrifices have been the foundation of my life's journey, and I am profoundly thankful for everything you have done for me. Thank you mom and dad!

In conclusion, this Dissertation represents the culmination of years of hard work, dedication, and collaboration. It is a testament to the support and mentorship I have received along the way. While this acknowledgment section may capture some of the gratitude I feel, the debt of gratitude I owe to everyone who has been a part of this journey is immeasurable. Thank you all for your unwavering support.