1

Introduction

1.1 Industrial Batch Processes

Industrial procedures involving raw material and mechanical steps in the manufacture of a product are usually carried out on a very large scale and in different ways. They can be broadly classified as continuous or batch procedures. In both cases, the goal is to operate the process so that desired end-product specifications are met. The achievement of these specifications involves technical decisions at the process operational level, which are related to reaching defined values in a set of relevant process variables through control systems.

Batch processes play an important role in the production of high-added-value specialties in key economic sectors such as manufacturing industry, agriculture, and biotechnology. The manufacture of these products requires a more sophisticated production strategy than the one required in a continuous process. In general, batch processes can be defined as repetitive and finite-duration processes consisting of several actions performed in a sequential manner (see Figure 1.1). A specific mix of raw materials or reactants is initially charged into a vessel. Then, materials are processed for a finite duration with a time-varying control configuration. Optionally, intermediate reactants can be added to the vessel at a proper time. When this is the case, the process may be called a fed-batch process. Once raw materials have been converted into desired products, they are discharged from the vessel and the vessel is cleaned.

This complex type of process has been experiencing a renaissance in the last decades as products-on-demand and first-to-market strategies impel the need for flexible and specialized production methods.[13] Several reasons make batch processes, with their repetitive sequence of charge/process/discharge operations, preferable over continuous processes, where feeding, processing and product extraction are continuous over time. In practice, keeping steady-state conditions over prolonged periods of time is difficult in some processes. For instance, in the biotechnological and pharmaceutical industries that make use of microorganisms, keeping their genetic stability over time is a tough problem. Keeping sterile conditions over prolonged periods of time is another typical problem in many biochemical

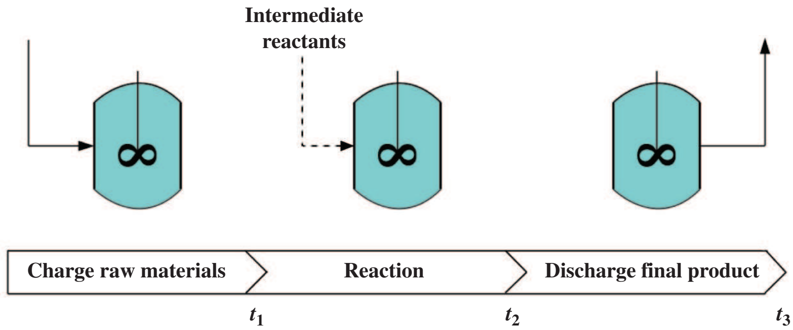

Figure 1.1 Batch process scheme representing the main stages for production.

processes. Also, frequent tuning of process settings for low-volume productions may be economically unfeasible in many processes. As a result, batch processing is usually preferred for low production volumes below 10,000 Mt/year while continuous ones are predominant for production volumes higher than 100,000 Mt/year in industries such as petrochemical, chemical, polymer, and food.[113]

For process control and monitoring purposes, a set of measurements belonging to different process variables is recorded in a batch process. Prior knowledge about the process may determine the sampling frequency to measure the variables, considering for instance the type of variable or the importance of certain phases in the process. Additionally, the quality of a batch can be quantified by collecting online quality measurements provided by probes, as is the case with ammonia probes in wastewater treatment processes. Another possibility to assess the quality of a batch is by analyzing a sample of the final product in a laboratory upon batch completion. These measurements can be seen as end-product quality indices that represent how well the batch was operated. As a consequence of the high cost to obtain these measurements, the frequency at which they are sampled is lower than for engineering measurements (e.g., temperatures, flows, levels). For this reason, and considering that results from the laboratory may be available after some delay, problems raised in a batch may be reproduced in subsequent batches before they are detected and properly diagnosed.

Advances in technology have permitted access to vast databases of engineering data collected in batch processes, which can be leveraged for process monitoring, troubleshooting and optimization. Batches that approximately follow a reference evolution and yield a product within specifications are referred to be under normal operating conditions (NOC). If this NOC can be somehow established, using historical data, as reference trajectories for the process variables, batches which behave abnormally by deviating from these trajectories, so that there is a potential risk of producing out-of-specification product, can be detected long before the quality variables are actually measured. This monitoring capability helps improve profitability in production.

In this book, we introduce a set of methodologies to make the most of the historical data collected in a batch process to improve its future performance. The methodologies are non-intrusive, a key advantage for acceptance in an industrial scenario, while revenues may be substantial. While not all methodologies will be suitable for the process at hand, we will present strategies for application that will help the implementation by the practitioner in his/her specific set-up. Numerous practical examples are included throughout the book to facilitate this goal.

This chapter describes the state-of-the-art in batch process modeling and monitoring. Section 1.2 introduces common types of sensors that can extract information from an industrial process. Section 1.3 briefly describes the major paradigms for modeling batch processes. Finally, in Section 1.4, the modeling cycle of batch processes is reviewed and discussed.

1.2 Types of Sensors

A sensor is any device that reacts to the environment in which it is placed and whose reaction is used to measure the property being sensed. There are many types of sensors to collect data in an industrial process. There are also many versions of each sensor type which may use a different sensing principle or may be designed to operate within different ranges. Some of them can measure physical properties such as temperature, pressure, level, humidity, speed, motion, distance, flow, etc. Others are designed to measure chemicals properties such as pH, oxidation–reduction potential, and concentration or amount of a given chemical species. In bioprocesses, a common type of sensor is a biosensor, which detects specific analytes thanks to a biological component, such as cells, proteins, nucleic acid or biomimetic polymers.

In order to describe and characterize the performance of a sensor, different indices have been proposed[38]:

- **Accuracy** is the degree of correctness when the device is operated under specified conditions. This is typically described in terms of a maximum percentage of expected deviation.
- **Limit of detection** is the smallest measurable input a sensor can signal, e.g., the lowest temperature measured by a thermometer. This concept differs from resolution, which defines the smallest measurable change in input that will produce a small but noticeable change in the output.
- **Linearity** is the degree to which the calibration curve of a device conforms to a straight line.
- **Precision** is the variability among several independent measurements of the same property under the same conditions.
- **Range** is the difference between the minimum and maximum values of sensor output in the intended operating window.
- **Reliability** indicates how well a sensor maintains both precision and accuracy over its expected lifetime.
- **Response time** is the time it takes for the sensor's output to reach its final value. This is a measure of the reaction time to changes in the environment, and must be compared with process dynamics.

- **Selectivity** is the ability of a sensor to measure only one metric or, in the case of a chemical sensor, to measure only a single chemical species.
- **Sensitivity** is the amount of change in a sensor's output in response to a change at the input over the sensor entire range. Provides an indication of a sensor's ability to detect changes. For some sensors, the sensitivity is defined as the input parameter change required to produce a standardized output change.

Some sensors provide a single number for each sample analyzed. In other cases, the provided response is a functional response (profile or spectrum), i.e., a function of some other magnitude such as wavelength (spectrometers) or time (electronic noses and tongues). In these cases, spectra have to be preprocessed (corrected) prior to data analysis in order to remove unwanted systematic variation such as baseline drift, multiplicative scatter effects and profile regions of low information content. Common approaches for preprocessing of spectral data are multiplicative signal correction, standard normal variate correction, Savitzky–Golay smoothing, and first and second order derivatives.[111]

Depending on the way the process sample is analyzed, sensor measurements can be classified as[121]:

- **Offline**, when a sample is analyzed in a laboratory.
- **At-line**, in the production area, during production close to the manufacturing process.
- **Online**, where the measurement system is connected to the process via a diverted sample stream; the sample may be returned to the process stream after measurement.
- **Inline**, where process stream may be disturbed (e.g., probe insertion), and measurement is done in real-time.
- **Noninvasive**, when the sensor is not in contact with the material (e.g., Raman spectroscopy through a window) in the processor, the process stream is not disturbed.

Offline sensors usually provide rich information on chemical properties by using highly specialized analytical instruments such as mass spectrometry, high-performance liquid chromatography or gas chromatography. The main disadvantages of offline sensors are the low-frequency rate due to the analysis cost, the high qualification required to operate the instrument, and the delay in providing the results due to the speed of the instrument and/or the preliminary treatment of the sample.

On the contrary, at-line, online, inline, and noninvasive sensors may provide process information at a fast rate with practically no delays and with simple measurement procedures. Inline sensors may suffer from contamination (e.g., in bioprocesses), leakage through the connections, and build-up of material on them, etc., corrupting the sensor output.

Advancements in biochemistry, chemistry, engineering, and materials science have been important drivers in the development of sensor technologies. Ongoing developments are offering possibilities for new sensors with advanced features, such as greater accuracy, lower cost, and increased reliability.

1.3 Batch Process Modeling

Two main batch modeling paradigms exist in the literature[8]: the mechanistic (or knowledge-based) models inherited from the field of classical mechanics, and data-driven models originally proposed in the field of engineering. The former are developed from the first-principle knowledge of the process in the form of mathematical or biochemical relations, such as reaction kinetics, stoichiometries, and mass/energy balances, that capture the behavior of the process. In contrast, the data-driven models are calibrated solely from historical data. The final goal is to build models that best describe the nature or sources of variation in the process over time. The choice of the model depends on the type of processes and the model purpose: exploratory analysis, optimization, monitoring, control, etc.[72]

The complexity of industrial processes has fostered the development of a large number of model structures and modeling methodologies spanning the whole range between both paradigms, which can be classified into three categories[14]: knowledge-based models, data-driven models, and hybrid models. These models aim to decompose the process data into different sources of variation:

1. A hard part, related to the deterministic behavior or prior process knowledge, such as mathematical relations describing chemical or physical properties.
2. A soft part, formed by the systematic variation that has not been taken into account in the hard part of the model, usually related to unknown sources of variation.
3. A residual part, expected to capture the unsystematic variation of the process, i.e., part of the data variation not explained by the previous parts.

Knowledge-based models decompose process variation into a hard and residual part; data-driven models decompose it into soft and residual parts; and hybrid models include the three parts.

The separation of the different types of variation in process modeling is crucial for process understanding, optimization, monitoring and prediction. Knowledge-based models do not necessarily fit process data and explain batch-to-batch variation as well as data-driven models do. Hence, the use of data-driven models that are capable of capturing not only the known but also the unknown variation sources is highly recommended. However, the use of hybrid models is a good alternative to data-driven models if process knowledge is available. These models use the prior information (hard part) to separate the major variation in data that is mainly generated by known physico-chemical interactions. It permits discovering new phenomena within data that are less abundant (soft part). The remaining variability is kept in the residuals, and is subjected to be modeled separately.

1.3.1 Knowledge-based Models

In this category, both the mathematical models obtained from physicochemical laws – first-principles or fundamental models – and those obtained from the expert knowledge of a process – expert systems – are included. The application of these modeling strategies to batch processes has received a lot of attention by the scientific community for many years, both using dynamical models based on state formulations[51] and rule-based expert systems.[79, 91] These models allow us to estimate the underlying theoretical states of a complex process, such as the growth of bacteria in a certain environment. Take the emergent field of systems biology as an example. In that area, the development of kinetic models of cellular metabolic networks based on quantitative experiments becomes a challenging and primary task for the biotechnological industry. Such models are used to understand the principles that govern cellular behavior and to achieve a predictive understanding of cellular functions for subsequent modeling of fermentations.[69, 156]

Nonetheless, knowledge-based models are seldom suitable for tasks related to monitoring or prediction in batch processes. This is due to the fact that such models are based on assumptions, which may not be accurate for the majority of the process states. Additionally, some unknown reactions may exist. Hence, the uncertainty caused by the unknown knowledge not taken into account in the models may lead to inaccuracies in fault detection and predictions.

1.3.2 Data-driven Models

Current industrial processes, both continuous and batch processes, collect an enormous amount of measurements belonging to hundreds or even thousands of process variables. The analysis of such data in order to distinguish between in-control and out-of-control situations becomes a challenging task. Often, such data are highly correlated (process variables are collinear) with a low signal-to-noise ratio and missing measurements.[95] In this environment, methods to extract the information that explains the majority of the variability of process data are required.[47, 98, 135] Nonetheless, batch data present a number of features which make this extraction cumbersome[166]:

- High dimensionality
- Nonlinear and time-varying dynamics
- Unequalized, uneven, and unsynchronized batch trajectories
- Presence of noise, collinear data, and outliers
- Variables of different magnitude and variance
- Missing data

Empirical multivariate models based on latent variables, such as principal component analysis (PCA) or partial least squares (PLS), coupled with other data processing methodologies developed in the literature, are useful to overcome the above problems, while also providing an analysis tool that is easy to understand from a statistical point of view. Other approaches have been proposed using support vector machines.[122] In batch multivariate statistical process control, a calibration data set composed of NOC batches that are representative of the common process variation is used for model building. Afterward, a test data set is typically used to validate the model parameters for subsequent use in process optimization, monitoring or prediction. For its multiple advantages and general applicability, this is the modeling approach that we will follow throughout the present book.

1.3.3 Hybrid Models

The use of knowledge-based models has the drawback that the unknown part of the process is not represented. Furthermore, some of the underlying assumptions (e.g., reaction kinetics, unknown dynamics, values of the model parameters and objective functions) may not be valid for all the possible states of the process.[89, 110] To overcome this problem, hybrid models combine knowledge-based models, which fit the theoretical behavior, and empirical models, which fit any remaining systematic variation.[48, 142]

Different approaches to decompose the data into the three types of variation (hard part, soft part and residuals) have been proposed in the literature. There are three categories in which these methods can be classified:

- Based on known constraints. There exist general frameworks that enable the imposition of very specific constraints on each type of information, i.e., a priori first principles information, or observed experimental information.[158] This can be done even at different temporal stages of the batch.[157] For batch processes, the basic methodology consists of using the a priori knowledge of a process for building a data-driven model by imposing different constraints, and then modeling the rest of the data using another model, in order to find those unknown sources of systematic variation. Interesting examples can be found in the literature.[72, 127, 170] More recently, grey component analysis (GCA) uses a soft penalty approach to gently force the empirical model toward the direction of the prior information – a chemically or biologically meaningful solution.[178] In summary, these models force the components to be a mixture between the data and the prior information.
- Focused on model parameter estimation. Another option is to use techniques based on introducing the prior knowledge by means of mathematical relations that describe the system behavior or dynamics. The starting point is some specific structure based on first principles, where some functions have to be estimated from data. Different tools can be used to approximate these functions, such as artificial neural networks (NN)[116] or Kalman filters.[11, 151]
- Constraints on the algorithms. Knowledge can also be incorporated in a model by imposing constraints on the modeling algorithms. For instance, some model parameters can be forced to have values within certain ranges, such as in multivariate curve resolution (MCR) or parallel factor (PARAFAC) models.[19, 159]

1.4 Bilinear Modeling Cycle for Batch Process Monitoring

Batch processes often exhibit batch-to-batch variations that are subject to investigation, analysis and monitoring: differences in the chemical composition, types and levels of impurities in raw materials, gradual operational changes, disturbances in the normal processing that affect the quality of the final product, etc. The analysis of data available in each batch is crucial to ensure safe operation, stable product quality and sustainable profit in batch processes. There are four major objectives for analyzing batch data[187]:

- **i.** The analysis of the variable trajectories of historical batches to gain process understanding and troubleshooting past abnormal operating conditions,
- **ii.** The statistical process control of incoming batches either after completion (so-called post-batch process monitoring) or during their progress (real-time process monitoring),
- **iii.** The prediction of the final product quality while the process evolves in real-time, and
- **iv.** The optimization and/or active control of batch operating conditions to reach the desired quality properties of the final product.

The available data in batch processes for statistical analysis mainly fall into three categories[43]:

- **i.** L initial conditions available before the start of the batch for I batches, which are arranged into a two-way array $\mathbf{Z}$ $(I \times L)$. These conditions include raw material properties, preprocessing times such as stage duration and waiting times, information on the shifts, any process measurements taken before the batch starts and expected to have an influence on quality, etc.
- **ii.** J time-varying variable trajectories and/or online analytical sensor responses measured at K different sampling times for all the I batches, which are arranged into the three-way array $\underline{\mathbf{X}}$ $(I \times J \times K)$. Note that under varying sampling rate of variables and/or batches, a preprocessing step of the data will be required to build this three-way array.
- **iii.** M quality and productivity measurements archived after batch completion in the two-way array $\mathbf{Y}$ $(I \times M)$. These quality properties can be collected at K_y different sampling times over the production of each batch, leading to the three-way array $\underline{\mathbf{Y}}$ $(I \times M \times K_y)$ though.

In the design of monitoring schemes, two phases are involved[47]: model building (phase I) and model exploitation (phase II). In the former, understanding the nature of the effects of varying initial conditions and process operating trajectories on the performance of the batches and on the final product quality is pursued.[187] Thereafter, the gained understanding and the statistical models are used to isolate and diagnose past poor operating conditions and set up statistical process control (SPC) schemes for monitoring purposes in the second phase. In model building for process monitoring, a number of steps are typically performed, namely (i) data alignment, (ii) data preprocessing, (iii) transformation of the three-way array to one or several two-way arrays for bilinear batch modeling, and (iv) statistical analysis and monitoring system design. These steps are iteratively repeated, as illustrated in Figure 1.2, to refine the monitoring system of the process prior to model exploitation. The particularities of the processing steps are due to the type of statistical modeling methods used throughout this document, PCA and PLS, which can be considered the state-of-the-art in process modeling. For this reason, we first introduce these methods in Chapter 2 and then move on to detail each of the processing steps.

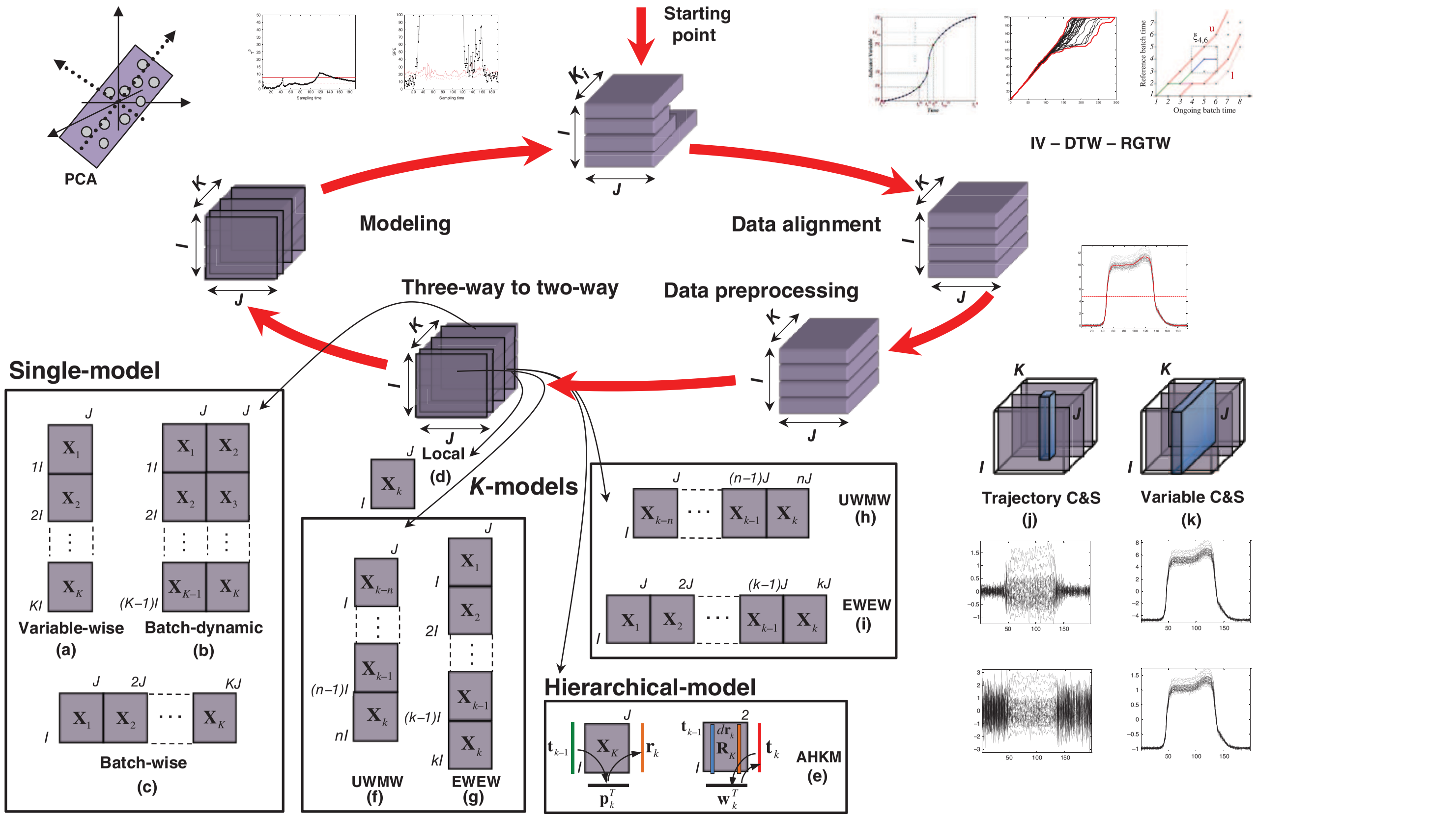

Figure 1.2 Modeling scheme in batch multivariate statistical process control systems based on PCA. *Source*: Reproduced from Ref.[64], with permission of John Wiley & Sons.

In a realistic set-up, the data collected from the process needs to be pretreated to yield data arrays $\mathbf{Z}$, $\underline{\mathbf{X}}$, and $\mathbf{Y}$. For instance, many of the variables may not be available for all batches and the sampling points for process and/or quality variables may not be the same for different batches or even different variables within the same batch. These mandatory preprocessing steps enable the configuration of data arrays, which is necessary for subsequent modeling and analysis. A fundamental problem is the data alignment of process variables to obtain a three-way data structure $\underline{\mathbf{X}}$ from the data collected through the network of process sensors with multiple sampling rates and for batches of possibly different duration and/or processing pace. This step involves two procedures: (i) batch data equalization and (ii) batch data synchronization. Data equalization involves homogenizing the data within a batch and across batches so that all values reflect process information at the same sampling time and frequency. Batch synchronization, on the other hand, aligns data from different batches to ensure that each sampling time represents similar process information at identical points in the process evolution across batches. The accuracy of empirical multivariate models and subsequent monitoring schemes in terms of fault detection and fault diagnosis is highly dependent on the quality of the equalization and synchronization.[64] This will be discussed in detail in Chapters 3 and 4.

After batch alignment and prior to model calibration, a preprocessing step is needed. Depending on the nature of batch data and the type of model to be fitted, the preprocessing approach may be different.[72] Again, a fundamental choice is the preprocessing applied for three-way matrices. Two main preprocessing methods are widely used in process chemometrics, although they might be referred with different terms: trajectory centering and scaling (trajectory C&S) and variable centering and scaling (variable C&S). The former consists of mean centering and scaling to unit variance the data corresponding to each jth process variable at each kth sampling time. Variable C&S performs mean-centering and scaling to unit variance of the whole data corresponding to each jth process variable. The discussion about which of these two choices is more adequate has been presented in the literature[73, 93, 180] since the two main pioneer research studies in batch monitoring[119, 186] selected one of them each. We address this topic in detail in Chapter 5.

After preprocessing, the transformation of the three-way into two-way data arrays for batch modeling is carried out. This is necessary in order to apply bilinear (two-way) modeling methods, such as PCA or PLS.[1] The different approaches to perform this transformation can be classified into three categories: (i) the single-model approach – e.g., batch-wise (BW), variable-wise (VW) and dynamic model, (ii) the K-models approach – e.g., uniformly weighted moving window (UWMW) and exponentially weighted evolving window (EWEW) models, and (iii) the hierarchical-model approach – e.g., adaptive hierarchical K-models (AHKM). In the single-model approach, the three-way array is unfolded into a single two-way array. The K-models approach is based on generating several bilinear models. The hierarchical-model approach is based on combining the past and current information at each sampling time in a hierarchical bilinear model. Chapter 6 discusses these different methods.

1 The statistical monitoring approach based on three-way modeling methods, with very little activity in the research community lately, is not treated in this book.

The final step of the bilinear modeling cycle of batch processes resorts to statistical analysis and the design of the monitoring system. This step is crucial to ensure safe and stable operation, and the production of a high-quality product meeting defined specifications. The statistical analysis and the design of the monitoring system is reviewed in Chapter 7.