44

PHARMACEUTICAL MANUFACTURING: THE ROLE OF MULTIVARIATE ANALYSIS IN DESIGN SPACE, CONTROL STRATEGY, PROCESS UNDERSTANDING, TROUBLESHOOTING, AND OPTIMIZATION

THEODORA KOURTI

Pharma Launch and Global Supply, GlaxoSmithKline, Global Functions

44.1 INTRODUCTION

The goal of any industry (be it chemical, pharmaceutical, steel, pulp, and paper) is to produce a product satisfactory to the customer (i.e., within prescribed quality specifications) under safety and environmental regulations and at a minimum cost. Quality control and regulatory specifications (safety, environmental) will help the manufacturer achieve the first three objectives but sometimes at the expense of cost. Understanding the process and monitoring and controlling process performance will help meet all four targets (quality product, safety constraints, environment constraints, minimum cost) simultaneously.

Process analysis and understanding, monitoring, and control have been practiced by several industries (notably petrochemical) for several decades. These industries adopted the above practices gradually. First, they saw the need for real-time quality measurements and developed real-time analyzers; as an example, the first analytical and control instrument group of UOP (Universal Oil Products) was formed in 1959 with the mission to develop online analyzers for internal pilot plant applications. This first step made the industry capable of collecting real-time measurements of quality properties and other process variables. The second step was the development of automatic process control techniques. This required some form of modeling. Attempts were made to understand the fundamental mechanisms of processes and built sophisticated mechanistic (first-principles) or empirical (data-driven) models. Later, in the 1990s the industry made a third step, which was the use of multivariate statistical analysis methods. With these multivariate approaches, it became possible to analyze and understand the process by looking at historical data containing hundreds of variables, detect abnormal situations, diagnose the sources of the abnormalities, and make appropriate modifications. Furthermore, by utilizing multivariate statistical process control (MSPC), it became possible to monitor the wellness of the process and product in real time, by looking simultaneously at hundreds of variables as they are collected. As a result, several industries managed not only to assure acceptable end-product quality, but also to improve process performance and maintenance, and to significantly reduce cost. The quick adoption of the methodologies and the benefits becomes evident from a very impressive set of applications presented by industry in 2003 in the symposium of "abnormal situation detection and projection methods—industrial applications [1]."

However, the picture of the pharmaceutical industry was different. At 2003, an article in the Wall Street Journal [2] proclaimed that "The pharmaceutical industry has a little secret: Even as it invents futuristic new drugs, its manufacturing techniques lag far behind those of potato-chip and laundry-soap makers." The article went on to explain that "in other industries, manufacturers constantly fiddle with their production lines to find improvements" but "regulations leave drug-manufacturing processes virtually frozen in

Chemical Engineering in the Pharmaceutical Industry: R&D to Manufacturing Edited by David J. am Ende

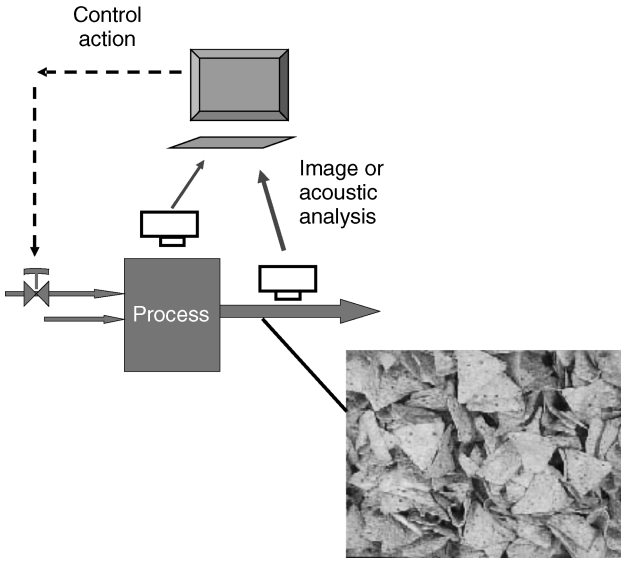

FIGURE 44.1 Monitoring and feedback control based on image and vibrational analysis. Example from snack food industry [3, 4].

time." Applications from the snack food industry (potato chip) that were published the same year [3, 4] were illustrating the use of inferential sensors based on a digital imaging system that had been developed for monitoring and control of the amount of coating applied to the base food product and the distribution of the coating among the individual product pieces, with real-time results from the implementation of such imaging system on snack food production lines. The imaging system was used to monitor product quality variables and to detect and diagnose operational problems in the plants. It was also used to implement closed loop feedback control over coating concentration. It was based on multivariate image analysis. Figure 44.1 shows the setup for feedback control based on Image Analysis. Images are collected by cameras from the process or from a stream exiting the process. This information then can be used in a feedback control loop.

The situation depicted in Figure 44.1 is the desired state in the pharmaceutical industry. That is, it is desired that we have the ability to measure or infer quality in real time, to assess the deviation from expected quality value, and to calculate a real-time control action for correction. Such ability stems from good models that provide process understanding, (different) models that convert spectral or other sensor data to quality, and (different) models that calculate the required control action. The word "different" was added on purpose in the previous phrase to indicate that in this endeavor different types and classes of models are employed and that there are several modeling activities required. The models required for all these classes of modeling may be first principles/mechanistic, data based/empirical, or hybrid. Multivariate projection methods or latent variable methods play an integral part in empirical and hybrid modeling. Such models can be developed to relate final quality properties to raw material attributes and process parameters and can be used for process understanding, process monitoring, and troubleshooting. Multivariate models can also be developed and utilized for process control, scale-up, and site transfer. There is a wealth of literature describing the theoretical foundation of the latent variable or projection methods [5–7], as well as the experiences from practitioners in industry [1, 8].

A lot of changes have happened since 2003 in the pharmaceutical industry that has entered a new era. The introduction of concepts such as quality by design, design space, and control strategy are also examples of such changes (Table 44.1). Multivariate methods are most suitable to address the requirements associated with these concepts. The pharmaceutical industry can learn from existing methodologies and from experiences from other industries and utilize multivariate technologies for fast process and product improvements.

In this chapter, the fundamentals behind latent variables modeling will be presented briefly together with references for in-depth presentations of methodologies and their use for process understanding, troubleshooting, monitoring, and control. Case studies are shown for such applications or are referenced. The chapter should be used by the reader as guidance for the types of problems that can be solved utilizing the methodology; the reader will use the detailed references to seek in-depth analysis and detailed solutions to specific problems.

TABLE 44.1 Terms Related to Quality by Design

Quality by design (*QbD*) is defined as a systematic approach to development that begins with predefined objectives and emphasizes product and process understanding and process control based on sound science and quality risk management
Design space is the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality
Control strategy is a planned set of controls derived from current product and process understanding that ensures (good) process performance and product quality. The controls can include parameters and attributes related to drug substance and drug product materials and components, facility and equipment operating conditions, in-process controls, finished product specifications, and the associated methods and frequency of monitoring and control

Note: These terms are defined by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH is a unique project that brings together the regulatory authorities of Europe, Japan, and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of product registration).

44.2 THE NATURE OF PROCESS AND QUALITY DATA

44.2.1 Multivariate Nature of Quality

Understanding the multivariate nature of quality is of great importance. Product quality is defined by the simultaneously correct values of all the measured properties; that is, product quality is a multivariate property. Most of the time, the property variables are not independent of one another, and none of them adequately defines product quality by itself; therefore, it is not a good practice to separately monitor key properties of the final product using univariate control charts.

Figure 44.2 is a classic illustration of the problem with using separate control charts for two quality variables (y_1, y_2). In this figure, the two variables are plotted against each other (upper left of the figure). The same observations are also plotted as individual (univariate) charts for y_1 (the horizontal plot) and y_2 (the vertical plot) with their corresponding upper and lower control limits. Suppose that when only common cause variation is present, y_1 and y_2 follow a multivariate normal distribution; the dots in the joint plot represent a set of observations from this distribution. Notice that y_1 and y_2 are correlated. The ellipse represents a $(1-\alpha)\%$ joint confidence limit of the distribution (i.e., when the process is in control, $\alpha\%$ of the points will fall outside the ellipse).

The point indicated by the $\oplus$ symbol is clearly outside the joint confidence region, and it is different from the normal in-control population of the product. However, neither of the univariate charts gives any indication of a problem for point $\oplus$; it is within limits in both of the charts. The individual univariate charts effectively create a joint acceptance region shaped like a square (shown with the ellipse). This will lead to accepting wrong products as good (point $\oplus$), but also rejecting a good product as bad (point $\diamondsuit$). The problem worsens as the number of variables increases. It is clear that an efficient fault detection scheme should look at the variables together.

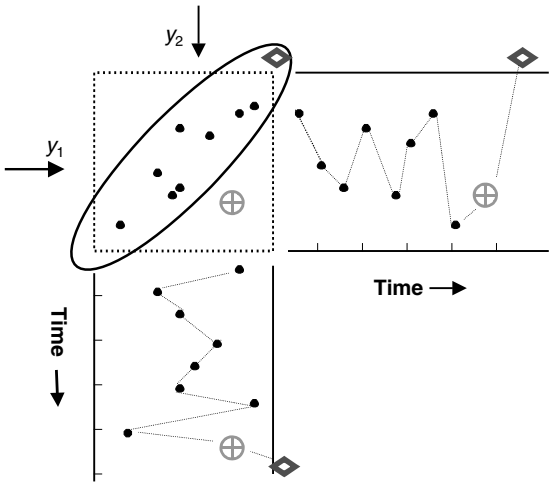

FIGURE 44.2 The multivariate nature of quality.

Multivariate charts are required to test quality [9] when it is described by many variables.

Recognizing the multivariate nature of quality should guide the procedures that will be used for the following cases:

- Raw material evaluation
- Intermediate quality evaluation
- Final quality evaluation
- Process control for quality
- Product transfer and scale-up (the multivariate nature of quality should be preserved for raw materials, intermediate qualities, and final qualities). This is a minimum requirement. Later, we will discuss the requirements on the multivariate space of the process variable trajectories as well.

44.2.2 Real-Time Monitoring and Process Signature: The Need to Utilize Information from Process Data

There is a widespread belief that the use of real-time quality measurements will help maintain the process "in control." Several practitioners are using real-time quality measurements to determine the "end point." The questions are the following:

- Does real-time "in-control" quality guarantee "in-control" process ?
- Is one or two final quality properties a sufficient "metric" of whether or not the process was in control?

Consider the example of Figure 44.3, where the trajectories of each one of the three process variables are plotted for three different batch runs, A, B, C. The final product at

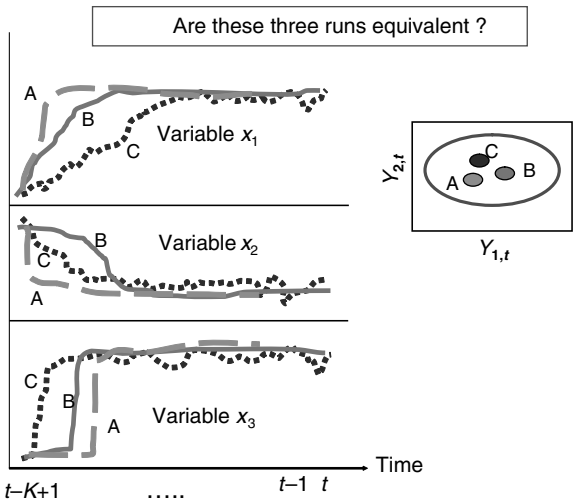

FIGURE 44.3 Different paths to the end point may result in similar values for some quality properties but may affect others in a different way. Consistent paths will assure that overall quality is consistent.

time t (end point quality) is determined from properties $y_{1,t}$ and $y_{2,t}$. These properties are plotted for each product produced by runs A, B, and C against the desired confidence limits of acceptable multivariate quality defined by the ellipse. Suppose that for all three runs the two "end properties" are on target in the multivariate control chart (points fall within the ellipse). However, the process trajectories follow different paths for each run. Are these runs equivalent?

It is well known in industry that the same measured quality properties can be achieved by taking different process paths. However, very frequently the few properties measured on a product are insufficient to define entirely the product quality. In polymer industry, for example, if only the viscosity of a polymer is measured and kept within specifications, any variation in end-use application (downstream processability) that arises due to variation in chemical structure (branching, composition, end-group concentration) will not be captured. To achieve consistency in all the product properties (measured quality and ability to process down the stream), the process conditions (path to end point) must also be kept in statistical control. When this is not the case, although the measured product properties may be on target, the properties that determined the processability of the product may not be within acceptable limits. Therefore, monitoring process data (temperatures, pressures, etc.,) together with real-time analyzers will give valuable information about events with special causes that may not only affect the final quality, but also give early warnings for potential equipment failure.

Another example that further corroborates this argument has been reported from the pharmaceutical industry [10]: "Conventional process control of drying of granulate in a fluidized bed drier would be to measure the loss on drying of a sample of powder, to determine water content. An advance on this may be to determine water content using an online NIR technology. However, true Process Understanding requires that the route by which you get to this end point be known and controlled. For example if the drying process is too vigorous, attrition may cause the granulate to generate an unacceptably high level of fine particles, which may cause downstream processing problems or dissolution issues; equally if drying is too slow, the potential for degradation of the drug molecule may exist."

This "process path to the end point" is also discussed in the European Regulatory Perspective [11], where it is reported that "during discussions within the industry, the term *process signature* has been mentioned regularly." To get a common understanding of this, the EU PAT Team had invited public comments on the following definition: "A collection of batch specific information that shows that a batch has been produced within a design space of the product." The EU PAT team mentions as examples of process signatures the amount of water added in relation to time (wet massing), air flow rate, and bed temperature during fall rate drying (fluidized bed drying). They concluded that their understanding is that there is no unique process signature, but instead a family of process signatures with common characteristics (salient features).

The above observations point to the importance of monitoring the process together with the product quality. By monitoring only the quality variables (in a univariate or multivariate chart), one performs statistical quality control (SQC). Real-time measurements on temperatures, pressure, pH, RPM, and so on combined with real-time measurements from analytical technology (spectroscopy, ultrasound, etc) will lead to online *process* monitoring and make MSPC, fault detection, and isolation possible. Combining information from the process measurements with the information from the analytical tools gives a very powerful tool to monitor the process. These two sets of measurements are not independent from one another but interrelated. As a matter of fact, these measurements "confirm" each other. This is the reason that process variables are sometimes used to assess the reliability of real-time analyzers. It will also be pointed out later that sometimes real-time process measurements may eliminate the need of some real-time analyzers. Information about the process may also include the vessels used for a specific run, the operators that were on shift, suppliers of raw material, and so on. Another advantage to using process measurements is that any abnormal events that occur will also have their fingerprints in the process data. Thus, once an abnormal situation is detected, it is easier to diagnose the source of the problem, as we are dealing directly with the process variables. For example, a pending equipment failure means that our production is not in control. However, there are situations that while there may be a pending equipment failure, real-time quality measurements may still be acceptable. By monitoring process variables, we have a very high probability to detect a pending problem.

Process data can be utilized together with appropriate models to

- Infer final product quality from process conditions during production
- Ease process understanding and troubleshooting
- Infer a quality in real time (soft sensors)
- Establish an overall "process signature" and monitor it
- Monitor analyzer reliability
- Check that the process is in a state of statistical process control (SPC)
- Decide on midcourse correction of variable trajectories to control final quality
- Establish operational knowledge that can be used for product transfer and scale-up

It should be emphasized here that although some relations between process operating conditions and final quality are known from the initial design of experiments (DOE), once we are in production, these relations may be influenced

TABLE 44.2 Some Multivariate Process Data Formats

Matrix Symbol	Dimensions	Explanation
$\mathbf{X}$	$(n \times k)$; two-way matrix; n observations in time, or n batches; k process variable measurements	Data from a continuous process, at given instant in time, or summary data from a batch (max T, min T, length of batch run, etc.)
$\mathbf{Y}$	$(n \times m)$; two-way matrix; n observations in time, or n batches; m product quality values	Quality data from a continuous process corresponding to the process measurements in $\mathbf{X}$, properly lagged, or quality data at the end of a batch.
$\underline{\mathbf{X}}$	$(I \times J \times K)$; three-way matrix; I batches; J process variables measured at K time intervals for each batch	Data collected from batch process at several time intervals during production.
$\mathbf{Z}$	$(n \times r)$; two-way matrix; n observations in time, or n batches; r other variable measurements	Raw material, total cycle times, length between processes, preprocessing information

by other factors and may change locally. By investigating production data, we can uncover the true relationships between process conditions and quality under the closed loop operations.

Finally, it is very frequently stated that critical process parameters should be identified and monitored together with critical quality attributes utilizing SPC. It should be emphasized here that the parameters that appear to be critical in the DOE are not necessarily the ones that will give information about the "wellness of the process" in SPC charts. The reason for this is that the parameters that are identified as important in DOE will be tightly controlled during production. SPC charts on routine data are noncausal; therefore, things that were important in DOE will not be important in SPC, unless something goes really wrong (i.e., the controller fails and cannot keep the desired target). As an example, suppose that temperature is important to the yield, as determined by DOE. This means that during production the desired temperature profile will be regulated by the controllers; in SPC monitoring, what is important (and in some types of processes will indicate the presence of excess impurities and other disturbances) is how much effort the controller is putting to maintain the temperature; that is, how much the valve to the cooling agent opened or closed during the reaction. Therefore, monitoring the controller action will provide much more information about abnormal situations than monitoring the temperature, although temperature was identified as critical process parameter by DOE.

44.2.3 Multivariate Nature, Structure, and Other Characteristics of Process Data

Databases containing measurements collected during production may become very large in size. The data are noncausal in nature (unless they come from designed experiments). They consist of highly correlated variables with many missing measurements and low content of information in any one variable (due to the low signal-to-noise ratios).

Multivariate Structure of Data: The convention that will be used throughout this chapter in expressing data is that of Table 44.2. Other formats that may appear in specific sections only will be defined in their corresponding sections.

44.2.4 Process Analysis and Process Understanding

Unfortunately, sometimes the term "process analysis" is wrongly being used as an equivalent term to process analytical chemistry. Collecting real-time measurements on a specific property may or may not reflect what is happening in the rest of process or the state of the process, unless the state of the process is completely observable in that quality, as discussed in Section 44.2.2; therefore, collecting a real-time quality measurement is not process analysis.

A process can be defined as a series of physical and/or chemical operations that converts input to output. Process analysis is a systematic examination of a process to understand it in order to develop ideas to improve it. Improvement could translate to better quality, lower cost, more efficient energy consumption, less pollutants to environment, and safer operation. One can perform process analysis utilizing both off-line and real-time measurements.

Process analysis leads to process understanding. Again, there may be several definitions of process understanding. There is a widespread belief that one gains process understanding only when they can describe the process by first principles, that is, by a theoretical or mechanistic model. However, one can gain tremendous insight into the process from empirical models derived from databases. These empirical models can lead to fast improvements that in several situations would have been impossible if people had been waiting for the development of theoretical first-principles models. Empirical models based on process data can be extremely valuable to diagnose abnormal operations such as pending equipment failure.

So while process understanding may by some definitions mean uncovering the mechanisms and path of a chemical reaction, or modeling a fermentation process, it may also

mean uncovering production problems such as the examples reported by practitioners in several industries:

- Diagnosing that the production of abnormal batches followed a specific pattern and as a result uncovering an incorrect operator practice that led to an abnormal batch every time a routine maintenance task was taking place.
- Understanding why recent process data projected on a latent variable space seem to form two clusters indicating different operation practices and hence solving an important operation problem as a result. Examination revealed that the cooling agent valve was not capable of meeting the demands in hot days and the reactor temperature could not be controlled properly. The valve was resized.
- Assessing that an operational problem caused the readings of a specific thermocouple to be erroneous and appear as outliers; thermocouple was too close to entrance of cool reactant; erroneous readings fed to the controller inappropriately alter the reactor temperature.
- Understanding where is the maximum process variability; is this variability noise or is it assignable to a cause—can we reduce it?

Understanding the way the process behaves in real scale production is a tremendous asset to the effort of product quality improvement and sometimes it weighs equally importantly to understanding the detail mechanisms of the reaction that takes place during production.

44.3 LATENT VARIABLE METHODS FOR TWO-WAY MATRICES

Latent variables exploit the main characteristic of process databases, namely, that although they consist of measurements on a large number of variables (hundreds), these variables are highly correlated and the effective dimension of the space in which they move is very small (usually less than 10 and often as low as 2). Typically, only a few process disturbances or independent process changes routinely occur, and the hundreds of measurements on the process variables are only different reflections of these few underlying events. For a historical process data set consisting of a $(n \times k)$ matrix of process variable measurements $\mathbf{X}$ and a corresponding $(n \times m)$ matrix of product quality data $\mathbf{Y}$, for linear spaces, latent variable models have the following common framework [12]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{44.1}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{44.2}$$

where $\mathbf{E}$ and $\mathbf{F}$ are error terms, $\mathbf{T}$ is an $(n \times A)$ matrix of latent variable scores, and $\mathbf{P}$ $(k \times A)$ and $\mathbf{Q}$ $(m \times A)$ are loading matrices that show how the latent variables are related to the original $\mathbf{X}$ and $\mathbf{Y}$ variables. The dimension A of the latent variable space is often quite small and determined by cross-validation or some other procedure.

Latent variable models assume that the data spaces $(\mathbf{X}, \mathbf{Y})$ are effectively of very low dimension (i.e., nonfull rank) and are observed with error. The dimension of the problem is reduced by these models through a projection of the high-dimensional $\mathbf{X}$ and $\mathbf{Y}$ spaces onto the low-dimensional latent variable space $\mathbf{T}$, which contains most of the important information. By working in this low-dimensional space of the latent variables $(t_1, t_2, \ldots, t_A)$, the problems of process analysis, monitoring, and optimization are greatly simplified. There are several latent variable methods. Principal component analysis (PCA) models only a single space ($\mathbf{X}$ or $\mathbf{Y}$) by finding the latent variables that explain the maximum variance. Principal components can then be used in regression (PCR). In PCR, there appears to be a misconception that the principal components (PC) with small eigenvalues will very rarely be of any use in regression. The author's personal experience is that these components can be as important as those with large variance. Projection to latent structures or partial least squares (PLS) maximizes the covariance of $\mathbf{X}$ and $\mathbf{Y}$ (i.e., the variance of $\mathbf{X}$ and $\mathbf{Y}$ explained, plus correlation between $\mathbf{X}$ and $\mathbf{Y}$). Reduced rank regression (RRR) maximizes the variance of $\mathbf{Y}$ and the correlation between $\mathbf{X}$ and $\mathbf{Y}$. Canonical variate analysis (CVA), or canonical correlation regression (CCR), maximizes only the correlation between $\mathbf{X}$ and $\mathbf{Y}$. A discussion of these latent variable models can be found elsewhere [12]. The choice of method depends on the objectives of the problem; however, all of them lead to a great reduction in the dimension of the problem. Some of them (PCR and PLS) model the variation both in the $\mathbf{X}$ space and in the $\mathbf{Y}$ space. This point is crucial in most of the applications related to PAT that are discussed in the following sections, as well as for the problem of treating missing data. The properties of PCA and PLS are discussed briefly below.

44.3.1 Principal Component Analysis

For a sample of mean centered and scaled measurements with n observations on k variables, $\mathbf{X}$, the principal components are derived as linear combinations $\mathbf{t}_i = \mathbf{X}\mathbf{p}_i$ in such a way that subject to $|\mathbf{p}_i| = 1$, the first PC has the maximum variance, the second PC has the next greatest variance and is subject to the condition that it is uncorrelated with (orthogonal to) the first PC, and so on. Up to k, PCs are similarly defined. The sample principal component loading vectors $\mathbf{p}_i$ are the eigenvectors of the covariance matrix of $\mathbf{X}$ (in practice, for mean centered data, the covariance matrix is estimated by $(n-1)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$). The corresponding eigenvalues give the variance of the PCs (i.e., $\mathrm{var}(\mathbf{t}_i) = \lambda_i$). In practice, one rarely needs to compute all k eigenvectors, since most of the

predictable variability in the data is captured in the first few PCs. By retaining only the first A PCs, the **X** matrix is approximated by equation 44.1.

44.3.2 Partial Least Squares

PLS can extract latent variables that explain the high variation in the process data, **X**, which is most predictive of the product quality data, **Y**. In the most common version of PLS, the first PLS latent variable $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ is the linear combination of the x variables that maximizes the covariance between $\mathbf{t}_1$ and the **Y** space. The first PLS weight vector $\mathbf{w}_1$ is the first eigenvector of the sample covariance matrix $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$. Once the scores for the first component have been computed, the columns of **X** are regressed on $\mathbf{t}_1$ to give a regression vector, $\mathbf{p}_1 = \mathbf{X}\mathbf{t}_1/\mathbf{t}_1^T\mathbf{t}_1$; the **X** matrix is then deflated (the $\hat{\mathbf{X}}$ values predicted by the model formed by $\mathbf{p}_1$, $\mathbf{t}_1$, and $\mathbf{w}_1$ are subtracted from the original **X** values) to give residuals $\mathbf{X}_2 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T$. **Q** are the loadings in the **Y** space. In the so called NIPALS algorithm, $\mathbf{q}_1$ is obtained by regressing $\mathbf{t}_1$ on **Y,** and then **Y** is deflated $\mathbf{Y}_2 = \mathbf{Y} - \mathbf{t}_1\mathbf{q}_1^T$. The second latent variable is then computed from the residuals as $\mathbf{t}_2 = \mathbf{X}_2\mathbf{w}_2$, where $\mathbf{w}_2$ is the first eigenvector of $\mathbf{X}_2^T\mathbf{Y}_2\mathbf{Y}_2^T\mathbf{X}_2$, and so on. The new latent vectors or scores ($\mathbf{t}_1$, $\mathbf{t}_2$, ...) and the weight vectors ($\mathbf{w}_1$, $\mathbf{w}_2$, ...) are orthogonal. The final models for **X** and **Y** are given by equations 44.1 and 44.2 [13, 14].

44.3.3 Latent Variables for Process Understanding

Latent variable methods are excellent tools for data exploration to identify periods of unusual/abnormal process behavior and to diagnose possible causes for such behavior (troubleshooting). The scores and loadings calculated by PCA and PLS and the weights by PLS can be utilized for this purpose. By plotting the latent variables (t_1, t_2, ..., t_A) against each other, the behavior of the original data set (be it process **X**, or quality data **Y**) can be observed on the projection space. By examining the behavior in the projection spaces, regions of stable operation, sudden changes, or slow process drifts may be readily observed. Outlier and cluster detection also becomes easy, both for the process and for the quality space. An interpretation of the process movements in this reduced space can be found by examining the loading vectors ($\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_A$) or ($\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_A$) in the case of PLS and the contribution plots. For a PCA analysis on **X** or a PLS analysis on **X** and **Y**, each point on a t_1 versus t_2 plot is the summary of measurements on k variables.

Figure 44.4a gives a simplified schematic interpretation of the methods. Suppose that we have measurements from five variables in a process (here, we plot the variable deviations from their nominal trajectories) during a time period. Suppose that variables x_1, x_3, and x_4 are correlated with each other, while variable x_2 is correlated with x_5. With the multivariate projection methods, new variables (latent variables) are calculated. In PCA, the first principal component t_1 is a weighted average of x_1, x_3, and x_4, while the second component, t_2, is a weighted average of x_2 and x_5. PCA can be seen as a classification of the main events that affect a process. The first principal component corresponds to the event that affects the largest number of variables, the second to the event that affects the next number of variables, and so on. The description here is a simplified explanation. It may be the case that two or more events affect the same variable, in which case this variable will contribute to the values of more than one component. We have reduced the number of the initial five raw variables to two principal components; we have *reduced the dimensionality* of the system. We can now plot these components against each other, as shown in Figure 44.4b. Each point on the plot summarizes the behavior of five raw variables. When the process is in statistical control, the points will be within the control limits, shown with an ellipse that is determined by statistical criteria (discussed later). If there is a problem in the

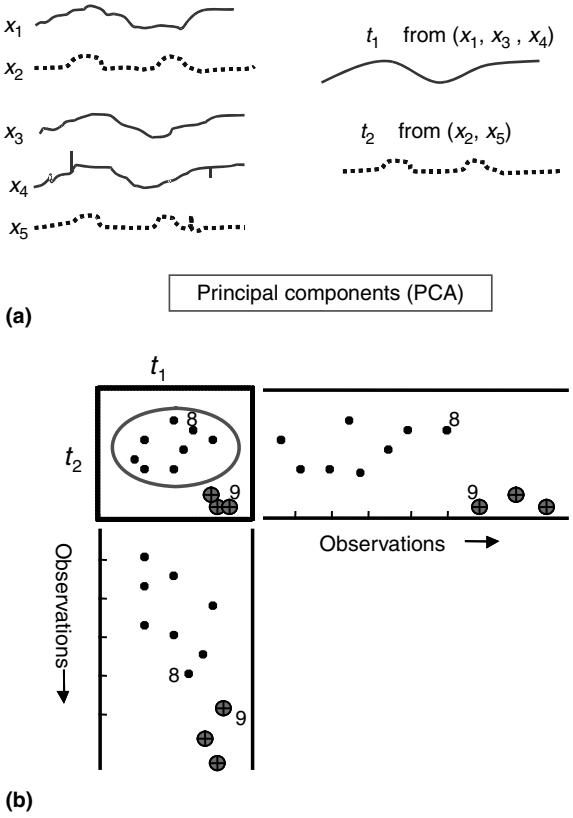

FIGURE 44.4 (a) Simple interpretation of PCA and dimensionality reduction. The principal components t_1 and t_2 use the correlation of five variables and break the process in two orthogonal events. The first principal component corresponds to the event that affects the largest number of variables, the second to the event that affects the next number of variables, and so on. (b) These components can be plotted against each other. A five-variable system is projected onto a two-dimensional plane.

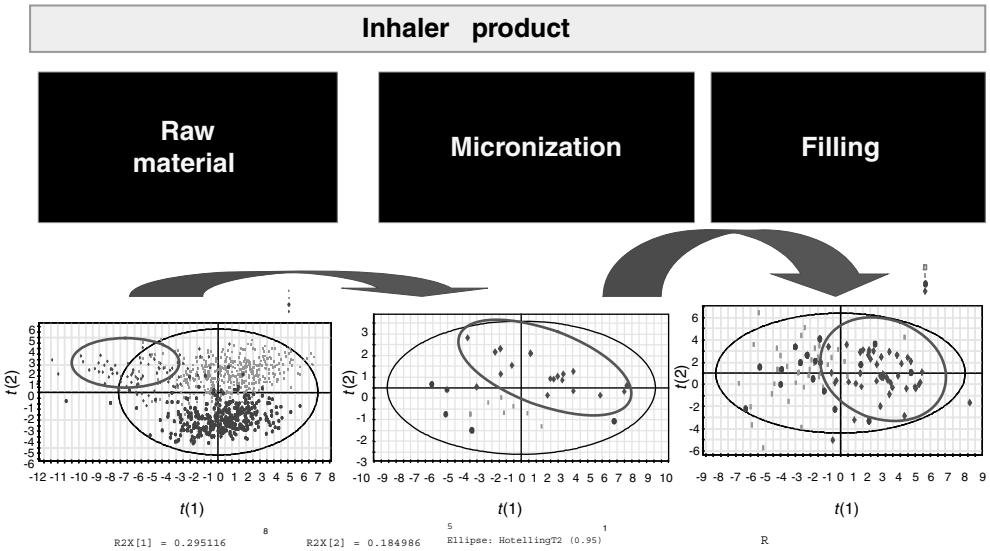

FIGURE 44.5 Representation utilizing projection methods. Raw material properties, micronization properties, and filling performance are projected on a latent variable space. Product batches produced from similar raw material (circled by the small ellipse) have similar filling performance.

process, the points will plot out of the ellipse. Notice that by using the latent variables, a five-variable system is *projected* onto a two-dimensional plane. This is why these methods are also called projection methods. Plotting principal components against each other is a good way to visualize process behavior and detect outliers and clusters. Typically, a small number of components are required to describe the main events in a unit (usually less than 10, and sometimes only 3 to 4).

44.3.3.1 Process Understanding Example 1: Relating Issues Across Unit Operations The power of projection methods in exploring large databases is demonstrated with the following example, shown in Figure 44.5, where we plot projections of the raw material quality, micronized material quality, and final quality. More than 25 measurements of physical and chemical properties are collected per lot of raw material. These variables are projected on a space defined by latent variables that allow us to visualize better the process behavior. In the particular example, raw material is produced at three supplier locations. The raw material properties are within univariate specifications at all locations. Projected on a multivariate space t_1 vs t_2, however, they form three clusters ; one cluster projects at the low part of the t_1 vs t_2 plot (negative values of t_2) while the other two clusters at the upper part of the plot (positive values of t_2). One of these two clusters is marked by a small ellipse. This indicates that in a multivariate sense the material possesses slightly different characteristics depending on the location it was produced (covariance structure changes with location). A few of the batches of this raw material were subsequently used for a specific product, and we show its behaviour after micronization and filling. The material properties after micronization are also projected on principal components and it could be observed that the material corresponding to the batches of the small ellipse projects on a different location from the rest of the batches. The filling performance of the material originating from the batches of the small ellipse is different from the rest of the material. The conclusion for this example is that the raw material differences propagate in the final quality. Given the large number of variables involved, it is clear that the projection methods provided a very quick diagnostic of the problem. This could have not been possible by dealing with univariate charts. The reader may note here that although the control ellipses shown are set by default in the vendor software, they are not interpretable when there is clustering; the assumptions for the calculation of these ellipses are for process monitoring and not for process exploration where there is intentional variation such as that introduced by design of experiments.

44.3.3.2 Process Understanding Example 2: Quick Diagnosis of Effect of Raw Material Variability to Granule Characteristics The advantage of PLS is that many highly correlated quality responses can be analyzed simultaneously. The **Y** matrix can contain several parameters related to quality. In this example, we wanted to see the correlation between certain API physical properties and the granule properties. We chose to use as **Y** the entire tapped density profile and the entire PSD profile. As seven variables corresponded to each property profile, we did not have to use any special scaling discussed in the multiblock section. First we use as **X** matrix the API properties and used PLS, between **Y** and **X** to relate API variability to the variability of the granule properties. Figure 44.6 shows the variability explained for each variable in the tapped density profile, and granule size profile, in a cross-validated model. Light grey is the direct

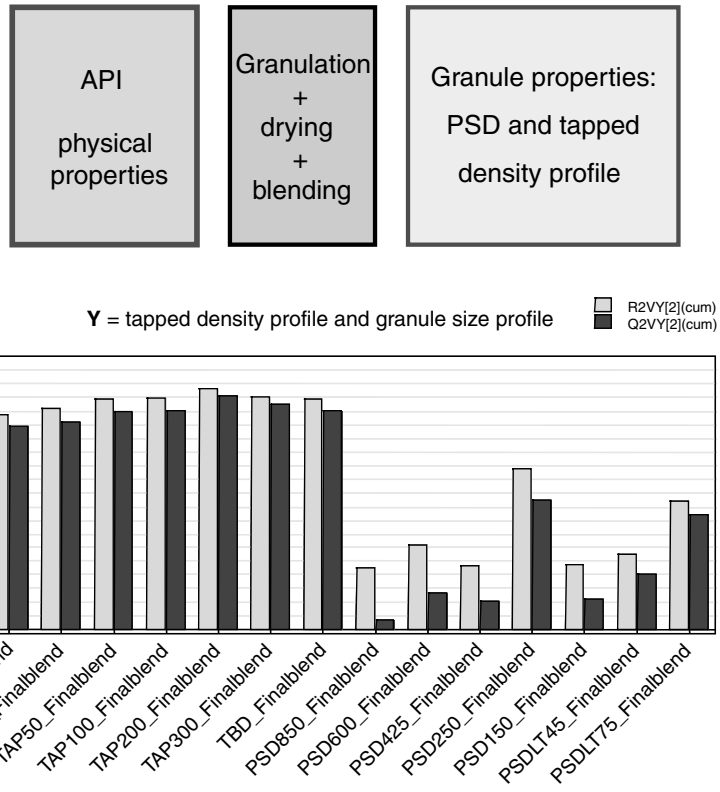

FIGURE 44.6 PLS is a powerful tool to help visualize multiple responses. Here, we study the effect of API physical property characteristics on the tap density profile and the granule size distribution.

fitting variability and dark the cross-validated variability. One can very clearly visualize that the tapped granule density is very highly correlated to the specific API physical property variability (as a high % of variability of the tapped density profile is explained by the variability of the API property values).

44.4 MULTIVARIATE STATISTICAL PROCESS CONTROL

From routine operation, we can establish acceptable limits of good process behavior. On a t_1 versus t_2 plane, such limits will take the form of an ellipse. When the process is in statistical control, the points will be within the ellipse. If there is a problem in the process, the points will plot out of the ellipse. In Figure 44.4b, the ellipse is calculated based on PCA on the data from good operation. Notice that while for raw correlated data the ellipse is tilted, indicating correlation (Figure 44.2), this is not the case when it is calculated for the principal components that are orthogonal.

To monitor the process in real time, however, it would have become cumbersome to have to plot all combinations of principal components (even if we had four components, we would need six charts). A statistic (Hotelling's T^2) can be calculated and the overall variability of the main events of the system can be monitored with a single chart, such as the one shown at the upper left corner of Figure 44.7. The line corresponds to acceptable performance. For the case of two components, this solid line corresponds to the perimeter of the ellipse of Figure 44.4b. For three components, it would correspond to the surface of an ellipsoid, and for four components the surface of a hyperellipsoid.

Hotelling's T^2 for scores is calculated as

$$T_A^2 = \sum_{i=1}^{A} \frac{t_i^2}{\lambda_i} = \sum_{i=1}^{A} \frac{t_i^2}{s_{t_i}^2} \tag{44.3}$$

where $s_{t_i}^2$ is the estimated variance of the corresponding latent variable t_i. This chart essentially checks if a new observation vector of measurements on k process variables projects on the hyperplane within the limits determined by the reference data.

As mentioned above, the A principal components explain the main variability of the system. The variability that cannot be explained forms the residuals (squared prediction error (SPE)). This residual variability is also monitored and a control limit for typical operation is being established. By monitoring the residuals (Figure 44.7, bottom left), we test that the unexplained disturbances of the system remain similar to the ones observed when we derived the model. For example, a model derived with data collected in the summer may not be valid in the winter when different

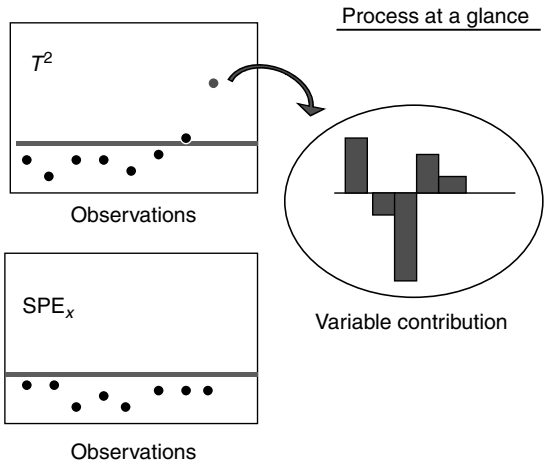

FIGURE 44.7 Two charts (T^2 and SPE) are required to give a picture of the wellness of the process at a glance. The hundreds of measurements collected from the process variables at each instant in real time are translated into one point for the T^2 chart and one point for the SPE chart.

disturbances affect the system (cooling water temperatures different, equipment walls colder, valves may reach limits in capacity of providing heating agent, etc). It is therefore important to check the validity of the model by checking the type of disturbances affecting the system. When the residual variability is out of limit, it is usually an indication that a new set of disturbances have entered the system; it is necessary to identify the reason for the deviation and it may become necessary to change the model.

SPE_X is calculated as

$$\text{SPE}_X = \sum_{i=1}^{k} (x_{\text{new},i} - \hat{x}_{\text{new},i})^2 \tag{44.4}$$

where $\hat{x}_{\text{new}}$ is computed from the reference PLS or PCA model. Notice that SPE_X is the sum over the squared elements of a row in matrix **E** in equation 44.1. This latter plot will detect the occurrence of any new events that cause the process to move away from the hyperplane defined by the reference model.

44.4.1 Calculation of Chart Control Limits

For a Hotelling's T^2 chart (either for PCA or PLS), an upper control limit based on the A first PCs and derived from n observations is obtained using the F distribution and given by

$$T^2_{A,\text{UCL}} = \frac{(n^2-1)A}{n(n-A)} F_{\alpha(A,n-A)} \tag{44.5}$$

where $F_{\alpha(A,n-A)}$ is the upper $100\alpha\%$ critical point of the F distribution with $(A,\ n-A)$ degrees of freedom.

For the SPE_X chart, limits can be computed using approximate results from the distribution of quadratic forms. The critical upper $100(1-\alpha)\%$ confidence interval on SPE is given as

$$\theta_1 \left[\frac{z_a \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{1/h_0} \tag{44.6}$$

where z_a is the unit normal deviate corresponding to the upper $100(1-\alpha)\%$, and α is the chance taken to incorrectly declare a fault because of the type I error,

$$\theta_i = \sum_{j=A+1}^{m} \lambda_j^i = Tr(\mathbf{E}^i) \quad \text{for } i = 1,\ 2,\ 3 \tag{44.7}$$

where λ_i is the ith eigenvalue referring to the covariance matrix and $h_0 = 1 - (2\theta_1\theta_3 / 3\theta_2^2)$.

Nomikos and MacGregor [15] used an approximation based on the weighted chi-square distribution ($g\chi^2(h)$). They suggested a simple and fast way to estimate the g and h that is based on matching moments between a $g\chi^2(h)$ distribution and the reference distribution of SPE at any time interval. The mean [$\mu = gh$] and the variance [$\sigma^2 = g^2(2h)$] of the distribution are equated to the sample mean (b) and variance (v) at each time interval. Therefore, g and h are estimated from the equations $\hat{g} = v/2b$ and $\hat{h} = 2b^2/v$.

Hence, the upper control limit on the SPE at significance level α is given by

$$\frac{v}{2b} \chi^2_\alpha \left(\frac{2b^2}{v} \right) \tag{44.8}$$

It should be emphasized that the models built for process monitoring model only common cause variation and not causal variation. The main concepts behind the development and use of these multivariate SPC charts based on latent variables for monitoring continuous processes were laid out in early 1990s [9].

These two charts (T^2 and SPE) are two complementary indices; together they give a picture of the wellness of the system at a glance. With this methodology, the hundreds of measurements collected from the process variables at each instant in real time are translated into one point for the T^2 chart and one point for the SPE chart (these two points summarize the process at that instant). As long as the points are within their respective limits, everything is in order. Once a point is detected out of limit, then the so-called *contribution plots* can be utilized that give us a list of all the *process* variables that mainly contribute to the out of limit point and hence allow us to diagnose the process problem immediately. Contribution plots can be derived for out of limit points in both charts.

Contributions to SPE: When an out of control situation is detected on the *SPE* plot, the contribution of each variable of the original data set is simply given by $(x_{\text{new},j} - \hat{x}_{\text{new},j})^2$. Variables with high contributions are investigated.

Contributions to Hotelling's T2: Contributions to an out of limit value in Hotelling's T^2 chart are obtained as follows: a bar plot of the normalized scores $(t_i/s_{ti})^2$ is plotted and scores with high normalized values are further investigated by calculating variable contributions. A variable contribution plot indicates how each variable involved in the calculation of that score contributes to it. The contribution of each variable of the original data set to the score of component q is given by the following equation:

$$c_j = p_{q,j}(x_j - \bar{x}_j) \text{ for PCA}$$

$$c_j = w_{q,j}(x_j - \bar{x}_j) \text{ for PLS} \qquad (44.9)$$

where c_j is the contribution of the jth variable at the given observation, $p_{q,j}$ is the loading, and $w_{q,j}$ is the weight of this variable to the score of the principal component q and $\bar{x}_j$ is its mean value (which is zero from mean centered data). Variables on this plot that appear to have not only the largest contributions to it, but also the same sign as the score should be investigated (contributions of the opposite sign will make the score only smaller). When there are K scores with high values, an "overall average contribution" per variable is calculated over all the K scores [6].

As an example, consider Figure 44.4b that illustrates that two clusters of points were observed on a t_1 versus t_2 plot. The use of contribution plots may help to investigate which variables have contributed to the move from point 8 to 9. So equation 44.9 would give the contribution of variable j to the move of the score values between two observations (say, 8 and 9) for component q calculated as [14]

$$p_{jq} \times (x_{j,9} - x_{j,8}) \text{ for PCA}$$

Utilizing contribution plots, when an abnormal situation is detected, the source of the problem can be diagnosed such that corrective action is taken. Some actions can be taken immediately, in real time. Others may require interventions to the process. One such example of an abnormal situation appeared in a reactor, in which the reactor temperature should be controlled in an exothermic reaction to 50°C. On a very hot day, the charts indicated abnormalities. Contribution plots pointed to a break in the correlation of cooling water flow and reactor temperature. It turned out that although the cooling water valve was fully open, it could not cope with the demand, as the cooling water was warmer. The valve had to be resized. MSPC pointed to a problem that had to be corrected. Therefore, the contribution plots are very important tools in understanding factors influencing the process during production and help in an "ongoing process understanding" philosophy.

44.4.2 How to Utilize the Control Charts

Since latent variable-based control charts were introduced, their use in industry is increasing. The charts answer the need of process industries for a tool that allows them to utilize the massive amounts of data being collected on hundreds of process variables, as well as the spectral data collected from modern analyzers.

Latent variable control charts can be constructed to monitor either a group of response variables **Y** (e.g., product quality variables) or a group of predictor variables **X** (process variables). For example, multivariate charts can be constructed to assess the consistency of the multivariate quality of raw materials, **Z**, and to test the final product **Y** for consistent quality. If there is spectral analysis on some of the materials, then multiblock concepts, discussed later, can be used.

A very important advantage of latent variables is that they can be used to monitor predictor variables taking into account their effect on the response variables. A model is built to relate **X** and **Y** using available historical or specially collected data. Monitoring charts are then constructed for future values of **X**. This approach means that the process performance can be monitored even at times when the product quality measurements, **Y**, are not available.

The main approach of SQC methods developed throughout the statistical literature has been to monitor only product quality data (**Y**) and, in some cases, a few key process variables (**X**). However, often hundreds of process variables are measured much more frequently (and usually more accurately) than the product quality data. So monitoring the process data is expected to supply much more information on the state of the process and supply this information more frequently. Furthermore, any special events that occur will also have their fingerprints in the process data. So, once a special event is detected, it is easier to diagnose the source of the problem as we are dealing directly with the process variables. On the contrary, control charts on the product variables only indicate that the product properties are no longer consistent with specification and they do not point to the process variables responsible for this.

Control charts on process variables are useful in multistep operations when quality data are not available between successive steps. For example, if a catalyst is conditioned in a batch process before being used for polymer production, the quality of the catalyst (success of conditioning) is assessed by its performance in the subsequent polymer production. It would be useful to know if the catalyst will produce good product before using it; monitoring the batch process variables with a latent variable chart would give early detection of poor quality product. Similarly, the few properties measured on a product are sometimes not sufficient to define product performance for several different customers. For example, if only viscosity of a polymer is measured, end-use applications that depend on chemical structure (e.g. branching, composition, end-group concentration) are unlikely to receive good material. In these cases, the process data may contain much more information about events

with special causes that affect the hidden product quality variables.

The philosophy applied in developing multivariate SPC procedures based on projection methods is the same as that used for the univariate or multivariate Shewhart charts. An appropriate reference set is chosen that defines the normal operating conditions for a particular process. Future values are compared against this set. A PCA or PLS model is built based on data collected from periods of plant operation when performance was good. Periods containing variations due to special events are omitted at this stage. The choice and quality of this reference set is critical to the successful application of the procedure.

44.5 BATCH PROCESS MONITORING

Figure 44.3 shows schematically the nature of batch process trajectories that are nonlinear and dynamic. Modeling batch operations requires taking into account their nonlinear dynamic nature. The methodology for developing multivariate control charts based on latent variables for batch process monitoring was initially presented by Nomikos and MacGregor [15–17] in a series of landmark papers. Any operation of finite duration, such as batch granulation, batch drying, blending/mixing additives for a finite time, fermentation, batch distillation, drying, and so on, can be modeled by the same methodology. This section will present the main issues that need be addressed in batch empirical modeling and will also give references to publications where these issues are discussed in detail.

44.5.1 Modeling of Batch Process Data

Most of the processes in the pharmaceutical industry are batch processes. Collecting real-time data during a batch process generates very large data sets. The top of Figure 44.8 gives the possible measurements that could be collected for a batch process. Information may be collected, at different time intervals for the duration of the batch, for several process variables such as agitation rate (RPM), pH, cooling agent flow (F), temperatures in different locations in the reactor (T_1, T_2, T_3). Data may also be collected in the form of spectra from real-time analyzers such as NIR. Finally, information may be available on raw material analysis, recipe, other preprocessing data, and even information on who was the operator on shift and which vessel was used.

Historical data collected from a batch process had traditionally been represented by a three-dimensional data array $\underline{\mathbf{X}}$ where a matrix $\underline{\mathbf{X}}$ ($I \times J \times K$) indicates that J process variables are measured at K time intervals, or K aligned observation numbers (A.O.N.), for each one of I batches.

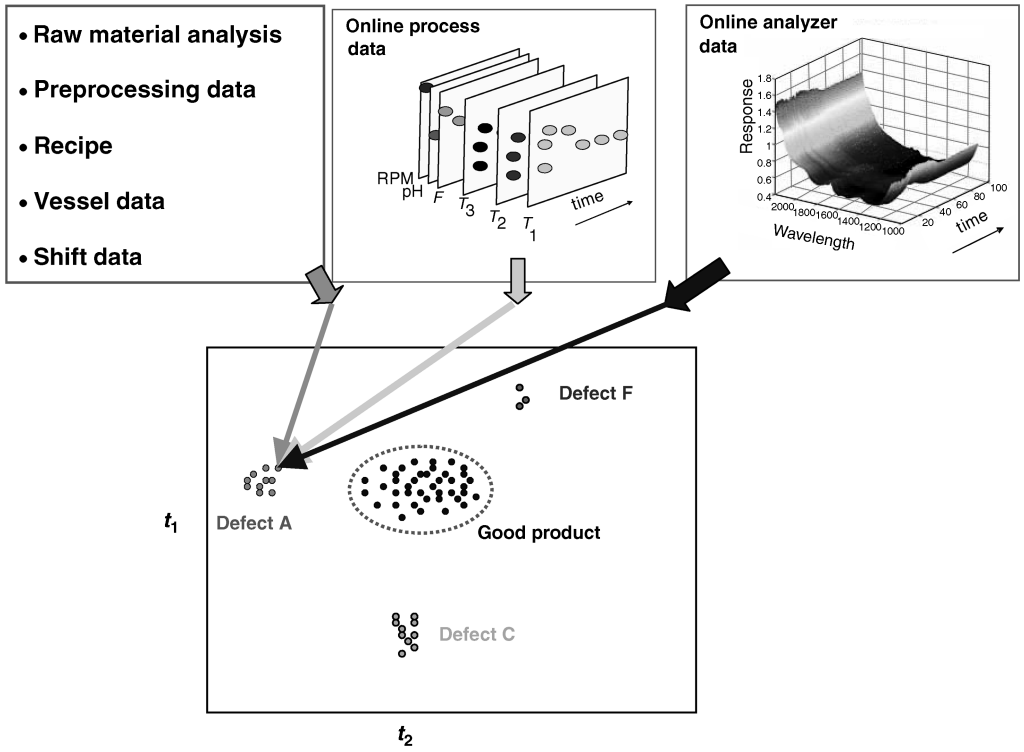

FIGURE 44.8 Data generated during batch runs are projected on a lower dimensional space, defined by two principal components t_1 and t_2. Each point on the plane corresponds to one batch run; that is, each point is the summary of the hundreds of measurements taken during the run.

Kourti [18,19] discussed that, in practice, it is not necessary that the same number of measurements are available for all the variables for the duration of the batch process. Some variables may not be present or measured for the full duration of the batch. Furthermore, the frequency of measurements may be different due to several reasons: (1) some variables may be measured more frequently than others (i.e., some every minute and others every 15 min); (2) certain phases in the process may be sampled more frequently to catch important phenomena (in emulsion polymerization, particle nucleation occurring at the very first few minutes in the reaction determines the number of particles and the particle size distribution; one may need to capture this with more frequent sampling at those stages). Therefore, Kourti argued the data set in such situations does not form a complete cube, but rather a cube where some columns are missing (Figure 44.9). Consequently, the methods used to model batch processes should be capable of modeling the structure of this incomplete cube. There are several methods for modeling three-way data. The choice of the method depends on the use of the model (i.e., prediction of final quality, monitoring, process control) and the types of the data sets available. Critical discussions on modeling procedures in batch processes for robust process monitoring, fault detection, and control can be found in selected publications [15–22].

The method presented by Nomikos and MacGregor [15] is termed in the literature as "batchwise unfolding" and is capable of modeling the incomplete cube structure. Furthermore, it is capable of modeling three-way structures generated when formulating the control problem of batch processes using latent variables that is discussed later. The method unfolds the three-dimensional structure into a two-dimensional array. In this new array, different time slices are arranged next to each other; variables observed at a given time interval are grouped in one time slice; the number of variables in each time slice may vary. Figure 44.9 shows an example of an unfolded matrix where variable x_4 is not measured at time $t = 1$, and variable x_2 is not measured at times $t = 3$–4. Once the three-way structure is unfolded to a two-way matrix **X,** equations 44.1 and 44.2 can be used to model **X** using PCA, or **X** and **Y** using PLS.

Multivariate control charts (Hotelling's T^2 and SPE) can be constructed for batch processes in a straightforward manner. Multivariate charts have superior detection capabilities to univariate charts for batch processes. In the words of a colleague from industry: "In most cases in practice, changes in the covariance structure precede detectable deviations from nominal trajectories. This was the problem that univariate monitoring approaches for batch processes could not address. In most process upsets it is the correlation among the monitored variables that changes first, and later, when the problem becomes more pronounced, the monitored variables deviate significantly from their nominal trajectories. There are cases where a process upset will change dramatically only the correlation among the variables without causing any of the variables involved to deviate significantly from its nominal trajectory. These particular cases, although rare, can result to significant cost to a company since they can go unnoticed for long periods of time (usually they are detected from a customer complaint)." (P. Nomikos, personal communication, 2002).

44.5.2 Alignment of Batch Processes of Different Time Duration

Sometimes batches have different time duration. In other words, using the same recipe, it may take different time to achieve the same conversion. This is due to the fact that time is not the deciding factor for the completion of the batch.

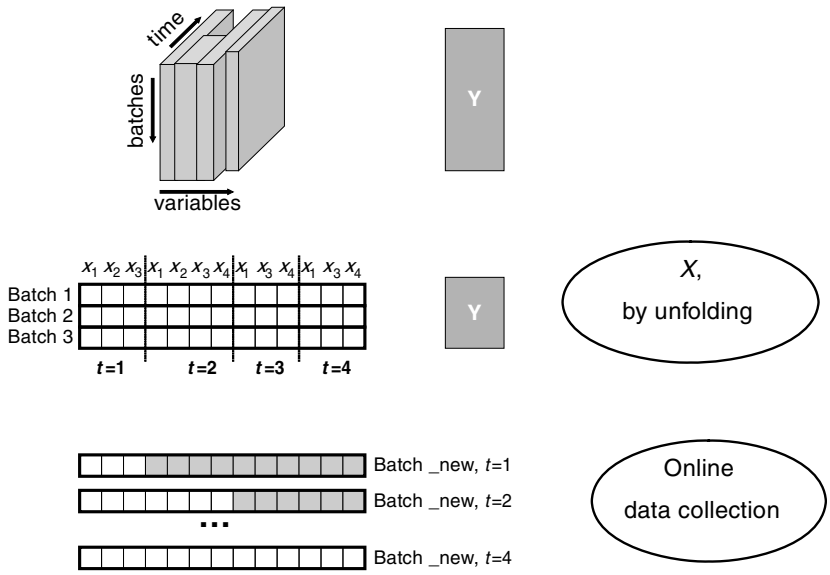

FIGURE 44.9 The structure of data in batch processes.

Sometimes the deciding factor may be the rate at which a certain reactant is added. One important stage before modeling batch process data is the alignment or synchronization of the data, such that they are expressed against the correct aligning factor (which may not be time). With alignment or synchronization, we must achieve the following: (1) Establish common start points at different phases of the run. For example, we could define that the first observation in all the runs for phase one of the reaction will correspond to the start of the monomer feed, for the second phase to the initiator injection, and so on; (2) match the shape of the trajectories of key variables. Once the shapes match, it is not necessary that the length of the batches match [19].

A critical discussion of the various synchronization approaches can be found in Kourti [19]. Attempts for batch data alignment involved the use of the cumulative amount of reactant added to the reactor as an indicator variable, where the variable trajectories were expressed as a function of the indicator variable, rather than time. The extent of the reaction was also used as an indicator variable later [23]. Dynamic time warping, based on speech recognition methods, was also suggested as well as the use of total time as a variable in the **Z** matrix, as extra information to describe the batch [24]. Taylor [25] suggested to include the cumulative warping, up to a given warped observation, as a new variable trajectory; his argument was that this would provide much richer information on the state of the batch by comparing it to the "typical batch" and would provide it in real time, rather than waiting for the batch to finish so that we can calculate the total time; the cumulative time spent could be used as extra trajectory in the case of alignment with an indicator variable. This suggestion was later used with excellent results and also provided the basis for designing batches with desired duration. The cumulative warp can be used as an extra variable to take into account time effects on batch quality when batches were synchronized by dynamic time warping, while García-Muñoz et al. [26] used the cumulative time when batches were synchronized by the indicator variable approach. Provided that an indicator variable exists (or can be constructed by nonlinear transformations from other variables and/or process knowledge), the indicator variable approach is usually chosen as the simplest and most convenient way for industrial applications.

44.5.3 Mean Centering and Scaling the Incomplete Cube

Mean centering the two-way matrix, formed by batchwise unfolding the three-way data, is equivalent to subtracting from each variable trajectory its average trajectory over the I batches, and thus converting a nonlinear problem to one that can be tackled with linear methods such as PCA and PLS.

When the three-way data form a full cube, that is an $\underline{\mathbf{X}}$ $(I \times J \times K)$ matrix, it is common practice to autoscale the two-way matrix formed by unfolding $\underline{\mathbf{X}}$ (i.e., divide each column by its standard deviation). This accomplishes two things: (1) gives an equal weight to all periods and consequently does not give high weights to noisy phases, or underweight a variable in tight control; after all, a variable is in tight control either because it is important to the product quality or because of safety and/or environmental concerns. (2) For the case of complete cube, it gives an equal weight to all the variables considered. However, in the case where variables are sampled less frequently, or are not present for the full run, the weights have to be adjusted accordingly, depending on the objective. To give equal weight to all variables, for example, after autoscaling, each column corresponding to variable j must be divided by $\sqrt{K_j}$, where K_j is the number of times that variable j was sampled in the run. In the example of Figure 44.9, after autoscaling the two-way matrix, all columns corresponding to x_1 and x_3 must be divided by $\sqrt{4}$, to x_2 by $\sqrt{2}$ and to x_4 by $\sqrt{3}$.

44.5.4 Online Monitoring of Batch Processes

Each batch run has a finite duration and the process variables exhibit a dynamic behavior during the run. This means that not only the autocorrelation structure of each variable changes during the run but also the cross-correlation of the variables changes. Models utilizing batchwise unfolding take into account this changing covariance structure of variables and time for the batch duration. For online monitoring, it is the structure of each evolving batch that it is compared against the typical behavior, as modeled by the training set of batches. The procedure for online monitoring of batch processes is slightly more complicated than that for continuous process because of the following reasons:

- For the online monitoring of continuous processes, at every time instant, we have a vector of new observations $\mathbf{x}_{\text{new}}^{\text{T}}$ that has a length equal to the number of columns in the model matrix **X** (and occasionally some measurements may be missing due to sensor failure, etc).
- In batch process monitoring, we have a vector with a length equal to the number of columns in the unfolded **X,** *only* when the batch run has finished. At any other time, data are missing from this vector, simply because they have not yet been collected. In Figure 44.9, the new vector batch_new is shown for different time intervals; gray areas have not been collected yet. Of course, the part with the collected data may also have missing data due to sensor failure, and so on.

Therefore, the score calculations and the limits for the multivariate control charts have to be developed in such a way that they take these "incomplete" measurement vectors into account. The procedure for the development of the multivariate control charts for the duration of the batch was

outlined by Nomikos and MacGregor [15]. To deal with the incomplete measurement vector, several approaches have been suggested. García-Muñoz et al. [27] recently investigated these approaches and demonstrated that using the *missing data* option and solving the score estimation problem with an appropriate method is equivalent to the use of an accurate forecast for the future samples over the shrinking horizon of the remainder of the batch. As PCA can model the covariance structure of the process variables, it facilitates handling of missing data. In batch processes, a PCA model describes the variance–covariance structure between variables over the entire batch; in other words, there exists information over all combinations of x_{jk}, j being the variable number and k the time interval (e.g., one can find how variable 4 at time 6 is related to itself at time 15 and also to variable 8 at time 35). Because of the tremendous structural information built into these multivariate PCA models for batch processes, the missing data option for predicting the future trajectory is shown to yield the best performance by all measures, even from the beginning of the batch.

Provided that there are no faults for the prediction of the future process variable trajectories, the final scores, and the product quality, these missing data estimation methods are very powerful. They have also been proven critical to the success of the control methods using latent variables. However, for process monitoring and online detection of process faults, all the alternative "filling in" methods give similar results. When a fault occurs, the model structure is not valid anymore. In that case, the differences among the trajectory estimation methods appear to be much less critical since the control charts used in each case are tailored to the filling in mechanism employed. All the approaches appear to provide powerful charting methods for monitoring the progress of batch processes.

The calculation of monitoring charts and their limits for batch processes, discussed in Ref. 15, where Hotelling's T^2 statistic for the analysis of batch process data (called D statistic) is calculated as

$$D = \mathbf{t}_R^{\mathrm{T}} \mathbf{S}^{-1} \mathbf{t}_R \mathbf{I} / (\mathbf{I} - 1)^2 \tag{44.10}$$

where $\mathbf{t}_R$ is the vector containing the R retained components of the model, and $\mathbf{S}$ represents the covariance matrix of the R retained score vectors. It is mentioned that $\mathbf{S}$ is a diagonal matrix due to the orthogonality of the scores, which is true for the final score estimate (i.e., when the batch run is complete). When computing this statistic for the online monitoring of batches, one should consider that the covariance of the scores changes with time and the scores might become nonorthogonal; therefore, Hotelling's statistic should be computed using the correct and complete variance–covariance matrix that corresponds to each time sample. This time-varying variance–covariance is computed using the reference set of batches. Therefore, the estimate of the Hotelling statistic at time k for batch i (D_{ki}) is a function of the estimate of the score vector for the R retained components at time k ($\hat{\tau}_{Rki}$) for batch i and the covariance matrix of the scores at time k ($\mathbf{S}_k$). Notice that D_{ki} will change depending on the method used to solve the missing data problem (or the option selected to "fill in"), since it is a function of $\hat{\tau}_{Rki}$ that has been shown [27] to differ from method to method and option to option. Using this corrected version of Hotelling's statistic dramatically improves abnormality detection.

44.5.5 Industrial Practice

Industrial applications for batch analysis, monitoring, and fault diagnosis have been reported [26–31]. It should be noted here that several companies choose to use the methodology not for real-time monitoring but as a tool to *real-time release* of the batch product in the following way: the batch is not monitored as it evolves, but rather immediately after the batch finishes, the process data are passed through the model and the scores for the complete batch are investigated. If they are within control limits, the product is released. If there is a problem, the product is sent for analysis in the laboratory. This procedure saves the company time and money. The batch run may last 2–3 h but the product analysis may take many more hours. That means that they do not have to waste batches while they are waiting for the results from the laboratory. By checking the process data as soon as the batch is complete, they can detect problems before starting a new batch. Other applications of multiway methods to batch analysis, optimization, and control have been reported and will be discussed later in the corresponding sections. Multiway methods and design of experiments can be used [32] to determine optimal process variable trajectories in a batch process in order to obtain a desired quality property.

There is a great potential for applications of multivariate batch analysis in the pharmaceutical industry. It is a superb tool for achieving process understanding. Using it to analyze past historical data will provide the user with summaries of the process history such as the one shown in Figure 44.8. The figure shows the projection of several batch runs on a plane, defined by two principal components t_1 and t_2. Each point corresponds to one batch run; that is, each point is the summary of the hundreds of measurements taken during the run. For example, for a batch run that lasts 16 h and where 5 min averages are collected on six process variable trajectories, each point is the summary of 1152 process measurements plus all the spectra scans at all time intervals plus other information on recipe, and so on. In the figure, one can immediately detect not only the cluster of good operation, but also clusters corresponding to specific product problems. Notice that the problems in product quality are observable by the projection of the process data and the measurements taken by real-time analyzers; that is, the problem in the

quality is observable without laboratory information. Utilizing contribution plots one can interrogate the multivariate model and determine combinations of variables and periods of operation that will drive a process away from producing a good product to producing defect A. Such an example is discussed in the literature [20] where information from the process and the raw materials was incorporated in the analysis. Based on this analysis, it was possible to determine raw material combinations and processing conditions that will result in a bad product.

An industrial example whereby monitoring the batch process variables a pending equipment failure was detected is reported in Ref. 31. A critical discussion on other batch processes modeling and monitoring procedures and other issues related to batch process analysis can be found in two recent studies [18, 19]. A discussion of the various approaches to synchronize runs of different durations can be found in Refs [13, 18, 19].

44.6 MULTISTAGE OPERATIONS: MULTIBLOCK ANALYSIS

There are multiple steps in pharmaceutical manufacturing and each step may involve multiple unit operations. Having a control chart for each unit rather than one for the whole process could be helpful to operators. However, building a model for each unit operation separately, does not consider interactions between unit operations. Such cases can also be addressed with latent variable models. Rather than building a model for each unit, one can build a model for the full process that will take into account the interactions between units and their relative importance to the final product quality by weighting them differently. Then, from this model, individual charts per unit operation can be derived. This way, interactions between unit operations are preserved. This is the approach of multiblock PLS (MB-PLS).

In the MB-PLS approach, large sets of process variables (**X**) are broken into meaningful blocks, with each block usually corresponding to a process unit or a section of a unit. MB-PLS is not simply a PLS between each block **X** and **Y**. The blocks are weighted in such a way that their combination is most predictive of **Y.** Several algorithms have been reported for multiblock modeling and for a good review, it is suggested that the reader consult Refs [33–35].

Multivariate monitoring charts for important subsections of the process, as well as for the entire process, can then be constructed, and contribution plots are used for fault diagnosis as before. In a multiblock analysis of a batch process, for example, one could have the combination of three blocks (**Z**, $\underline{\mathbf{X}}$, and **Y**); block **Z** could include information available on recipes, preprocessing times, hold times, as well as information of the shifts (which operator was in charge) or the vessels used (i.e. which reactor was utilized); $\underline{\mathbf{X}}$ would include process variable trajectories; and **Y** would be quality. Analysis of this type of data could even point to different ways the operators operate the units and relate product quality to operator, or different process behavior of vessels and identify faulty vessels, and so on. The reader is referred to the work of García-Muñoz et al. [20] for detailed examples where the multiblock analysis is utilized in batch processes for troubleshooting.

Several alternative ways to perform multiblock appear in commercial software. One approach that is being frequently used to deal with a data structure of several blocks involves two stages: PCA is performed for each one of the **Z** and **X** blocks and then the scores and/or residuals derived from these initial models are related to **Y** with a PLS. In an alternative version, PLS is performed between **Z** and **Y** and **X** and **Y,** and the resulting scores are related to **Y**. The users should exercise caution because these approaches may fail to take into account combinations of variables from different blocks that are most predictive of **Y**. For example, in situations where process parameters in **X** are modified to account for variability of raw material properties in **Z** (i.e., when **X** settings are calculated as a feedforward control to deviations of **Z**), a PLS between **Z** and **Y** will show that **Z** is not predictive of **Y** variability; similarly, a PLS between **X** and **Y** will show that **X** is not predictive of **Y**; a MB-PLS of **[Z, X]** and **Y** will identify the correct model. Finally, MB-PLS handles missing data in a very effective way.

As might be expected in multistage continuous processes, there can be significant time delays between the moment an event occurs in one unit (and therefore affects the variables of that unit) and the moment its effect will become obvious on a product variable at the end of the process. These delays significantly affect the interaction and correlation structures of the process variables and need to be handled by lagged variables created from the original process variables. Data can be time shifted to accommodate time delays between process units.

In some multistage operations, the path of the product through the various process units can be easily traced, and eventually one can relate a specific lot number to several process stages (via a multiblock PLS). In such cases, the process conditions of these units can be used to predict the quality of the product. There are situations, however, where a product (or the composition of the effluent stream of a process) is a result of a multistage operation but its path cannot be traced clearly due to mixing of streams from several parallel units in one vessel and then splitting to a number of other vessels. A discussion on monitoring difficult multistage operations can be found in Ref. 19. In those cases, the best alternative to achieve consistent operation is to monitor each unit, separately, by a PCA model. By assuring a consistent operation per unit, one hopes for a consistent product. Once an unusual event is detected in one unit, one may decide not to mix the product

further, or investigate lab quality before proceeding to the next stage.

44.7 PROCESS CONTROL TO ACHIEVE DESIRED PRODUCT QUALITY

The term "control" currently appears in the pharmaceutical literature to describe a variety of concepts, such as end point determination, feedback control, statistical process control, or simply monitoring. Process control refers to a system of measurements and actions within a process intended to ensure that the output of the process conforms with pertinent specifications.

In this chapter, we use some terms related to process control with the following definition:

- *Feedback Control*: to indicate that we are reactive; that is, the corrective action is taken on the process based on information from the process output (e.g., measurements on product quality, at given time.)
- *Feedforward Control*: to indicate that we are proactive; that is, the process conditions are adjusted based on measured deviations of the input to the process (e.g., information on raw material)

44.7.1 Feedforward Estimation of Process Conditions

The concept of adjusting the process conditions of a unit based on measured disturbances (feedforward control) is a concept well known to the process systems engineering community for several decades. The methodology is also used in multistep (multiunit) processes where the process conditions of a unit are adjusted based on information of the intermediate quality achieved by the previous unit (or based on raw material information).

An example of a feedforward control scheme in the pharmaceutical industry, where multivariate analysis was involved, is described by Westerhuis et al. [36] The authors related crushing strength, disintegration time, and ejection force of the tablets with process variables from both the wet granulation and tableting steps and the composition variables of the powder mixture. They also included physical properties of the intermediate granules. The granule properties may differ from batch to batch due to uncontrolled sources such as humidity temperature, and so on. This model is then used for each new granulation batch. A feedforward control scheme was devised that can adjust the variables of the tableting step of the process based on the intermediate properties to achieve desirable final properties of the tablets.

To the author's knowledge, there are several unpublished examples in the chemical and other industries where information on the raw data $\mathbf{Z}$ is used to determine the process conditions $\mathbf{X}$ or $\underline{\mathbf{X}}$ in order to achieve the desired quality $\mathbf{Y}$, utilizing projection methods. Sometimes such information from $\mathbf{Z}$ may simply be used to determine the length of the run, while in other cases, it may be a multivariate sophisticated scheme that determines a multivariate combination of trajectories for the manipulated variables. To achieve this, historical databases can be used to develop multiblock models $\mathbf{Z}$, $\mathbf{X}$ (or $\underline{\mathbf{X}}$), and $\mathbf{Y}$.

44.7.2 End Point Determination

There have been reports in the literature where real-time analyzers are used for "end point detection" or "end point control." In most of these situations, a desired target concentration is sought, for example, % moisture in drying operations.

An example is described by Findlay et al. [37], where NIR spectroscopy is used to determine granulation end point. The moisture content and the particle size determined by the near-infrared monitor correlate well with off-line moisture content and particle size measurements. Given a known formulation, with predefined parameters for peak moisture content, final moisture content, and final granule size, the near-infrared monitoring system can be used to control a fluidized bed granulation by determining when binder addition should be stopped and when drying of the granules is complete.

44.7.3 Multivariate Manipulation of Process Variables

It was discussed in Section 44.2.2 that regulating only the final value of a property (or even several properties) is not sufficient. In other words, end point control may not be sufficient. The process signatures are equally important. These process signatures should be regulated in a correct, multivariate way, not simply on a univariate basis. It is possible that two batch runs produce products with different quality, even if the trajectory (path to end point) of one quality variable follows the same desired path in both the runs. This will happen if the covariance structure of the trajectory of this variable with the trajectories of the rest of the process variables (temperatures, agitation rate, reactant addition) is different for these two batches. This concept is very important both in control and in scale-up. Latent variable methodology allows taking into consideration the process variable trajectories in a multivariate way.

Control of batch product quality requires the online adjustment of several manipulated variable trajectories. Traditional approaches based on detailed theoretical models are based on either nonlinear differential geometric control or online optimization. Many of the schemes suggested in the 'literature require substantial model knowledge or are computationally intensive and therefore difficult to implement in practice. Empirical modeling offers the advantage of easy model building.

Lately, latent variable methods have found their way to control batch product quality and have been applied in industrial problems. Latent variable methodology allows taking into consideration the process signatures in a multivariate way for end point detection problems. Marjanovic et al. [38] describe a preliminary investigation into the development of a real-time monitoring system for a batch process. The process shares many similarities with other batch processes in that cycle times can vary considerably, instrumentation is limited, and inefficient laboratory assays are required to determine the end point of each batch. The aim of the work conducted in this study was to develop a data-based system able to accurately identify the end point of the batch. This information can then be used to reduce the overall cycle time of the process. Novel approaches based upon multivariate statistical techniques are shown to provide a soft sensor able to estimate the product quality throughout the batch and a prediction model able to provide a long-term estimate of the likely cycle time. This system has been implemented online and initial results indicate that it offers the potential to reduce operating costs.

In another application [39], latent variable methodology was used for soft sensor development that could be used to provide fault detection and isolation capabilities and can be integrated within a standard model predictive control framework to regulate the growth of biomass within a fermenter. This model predictive controller is shown to provide its own monitoring capabilities that can be used to identify faults within the process and also within the controller itself. Finally, it is demonstrated that the performance of the controller can be maintained in the presence of fault conditions within the process.

Work has also been reported for complicated control problems where adjustments are required for the full manipulated variable trajectories [40]. Control through complete trajectory manipulation using empirical models is possible by controlling the process in the reduce space (scores) of a latent variable model rather than in the real space of the manipulated variables. Model inversion and trajectory reconstruction are achieved by exploiting the correlation structure in the manipulated variable trajectories. Novel multivariate empirical model predictive control strategy (LV-MPC) for trajectory tracking and disturbance rejection for batch processes, based on dynamic PCA models of the batch processes, has been presented. The method presented by Nomikos and MacGregor [15] is capable of modeling three-way structures generated when formulating the control problem of batch processes using latent variables.

44.7.4 Setting Raw Material Multivariate Specifications as a Means to Control Quality

Dushesne and MacGregor [41] presented a methodology for establishing multivariate specification regions on raw/incoming materials or components. The thought process here is that if the process remains fixed, we should control the incoming material variability. PLS is used to extract information from databases and to relate the properties of the raw materials supplied to the plant and the process variables at the plant to the quality measures of the product exiting the plant. The specification regions are multivariate in nature and are defined in the latent variable space of the PLS model. The authors emphasize that although it is usually assumed that the raw material quality can be assessed univariately, that is, by setting specification limits on each variable separately, this is valid only when the raw material properties of interest are independent of one another. However, most of the times the properties of products are highly correlated. In other words, treating the raw material properties in a univariate way, for two properties, it would mean that (referring to Figure 44.1) while we can process only material that falls in the ellipse, we agree to buy material from the supplier with the specifications set in the square; that is, we agree to use material that we know in advance it will not perform well.

To develop models to address the problem, multiblock PLS is used for $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{Y}$; $\mathbf{Z}$ contains measurements on N lots of raw material data from the past; $\mathbf{X}$ contains the steady-state processing conditions used to process each one of the N lots; $\mathbf{Y}$ contains final product quality for these N lots. The methodology could be easily extended to batch process $\underline{\mathbf{X}}$.

It should become one of the priorities in industries to express the raw material orders as a multivariate request to the supplier.

44.8 OTHER APPLICATIONS OF LATENT VARIABLE METHODS

44.8.1 Exploiting Databases for Causal Information

Recently, there has been a lot of interest in exploiting historical databases to derive empirical models (using tools such as Neural Networks regression or PLS) and use them for process optimization. The idea is to use already available data rather than collecting new through a design of experiments. The problem is that for process optimization causal information must be extracted from the data, so that a change in the operating variables can be made that will lead to a better quality product, or higher productivity and profit. However, databases obtained from routine operation contain mostly noncausal information. Inconsistent data, range of variables limited by control, noncausal relations, spurious relations due to feedback control, and dynamic relations are some of the problems the user will face using such happenstance data. These are discussed in detail in the section "Hazards of fitting regression equations to happenstance

data" in Ref. 42 where the advantage of experimental designs as a means of obtaining causal information is emphasized. In fact, in a humorous way, the authors warn the young scientists that they need a strong character to resist the suggestion of their boss to use data from past plant operation every time they suggest performing designed experiments to collect data.

In spite of this, several authors have proposed approaches to optimization and control based on interpolating historical bases. However, in all these cases, their success was based on making strong assumptions that allowed the database to be reorganized and causal information to be extracted. One approach was referred to as "similarity optimization" that combined multivariate statistical methods for reconstructing unmeasured disturbances with nearest neighbor methods for finding similar conditions with better performance. However, it too was shown to fail for many of the same reasons. In general, it was concluded that one can optimize only the process if there exist manipulated variables that change independently of the disturbances and if disturbances are piecewise constant, a situation that would be rare in historical process operations.

The reader should therefore exercise caution about how historical databases are used when it comes to retrieving causal information. However, databases obtained from routine operation are great a source of data for building monitoring schemes.

44.8.2 Product Design

Given the reservations about the use of historical databases, one area where some success has been achieved is in identifying a range of process operating conditions for a new grade of product with a desired set of quality properties and in matching two different production plants to produce the same grade of product. If fundamental models of the process exist, then these problems are easily handled as constrained optimization problems. If not, optimization procedures based on response surface methodology can be used. However, even before one performs experiments, there exists information within the historical database on past operating conditions for a range of existing product grades.

In this case, the historical data used are selected from different grades and therefore contain information on variables for several levels of past operation (i.e., there is intentional variation in them, and they are not happenstance data). The key element in this empirical model approach is the use of latent variable models that both reduce the space of $\mathbf{X}$ and $\mathbf{Y}$ to a lower dimensional orthogonal set of latent variables and provide a model for both $\mathbf{X}$ and $\mathbf{Y}$. This is essential in providing solutions that are consistent with past operating policies. In this sense, principal component regression and PLS are acceptable approaches, while MLR, neural networks, and reduced rank regression are not.

The major limitation of this approach is that one is restricted to finding solutions within the space and bounds of the process space $\mathbf{X}$ defined by previously produced grades. There may indeed be equivalent or better conditions in other regions where the process has never been operated before, and hence where no data exist. Fundamental models or more experimentation would be needed if one hopes to find such novel conditions.

A very good discussion on these issues can be found in García-Muñoz et al. [26]. The authors illustrate a methodology with an industrial batch emulsion polymerization process where the batch trajectories are designed to satisfy certain customer requirements in the final properties of the polymer while using the minimal amount of time for the batch run. The cumulative time, or used time, is added as an extra variable trajectory after the alignment of the batches.

44.8.3 Site Transfer and Scale-Up

Product transfer to different sites and scale-up falls into the same class of problems: one needs to estimate the process operating conditions of plant B to produce the same product that is currently produced in plant A.

Attempts have been made to solve such problems with latent variable methods, utilizing historical data from both locations for transferring other products.

The main points to keep in mind when addressing such a problem are as follows:

- The quality properties of the product should always be checked within a multivariate context because univariate charts may be deceiving. The multivariate quality space for both the sites should be the same. Correct product transfer cannot be achieved by comparing end point quality on univariate charts from the two sites (or from pilot scale and manufacturing). The product quality has to be mapped from site to site in a multivariate way (the products in both sites have to project on the same multivariate space).
- The end point quality may not be sufficient to characterize a product. The path to end product is important. Whenever full mechanistic models exist, these models describe the phenomena that are important for the process and therefore determine this path. When changing sites, the full mechanistic model will describe the desired path in the new site taking into account size, mass, and energy balances and/or other phenomena related to the process. When mechanistic models do not exist, this mapping of the "desired process paths" or "process signatures" has to happen with empirical data.

A methodology has been developed for product transfer and scale-up based on latent variables [43]. The methodology utilizes databases with information on previous products and their corresponding process conditions from both sites. The two sites may differ in equipment, number of process variables, locations of sensors, and history of products produced.

44.9 QUALITY BY DESIGN

44.9.1 Design Space: Expressing Quality as a Function of Input Material Attributes and Process Parameters

Several regulatory agencies participating in the International Conference of Harmonization have adopted the concept of design space. The definition of design space as "the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality" [44] stems from a well known fact that if the variability in the raw material is not compensated by the process, it will be transferred to quality.

The effect of the raw material attributes on the process performance, if the process operating conditions remain fixed, is clear in the example depicted in Figure 44.5. Recall that in that example the raw material is characterized by more than 25 physical and chemical properties and that these variables are projected on a space defined by latent variables that allow us to visualize better the process behavior. Raw material is produced at three supplier locations. Although the raw material properties are within univariate specifications, at all locations, the projection in three clusters indicates that in a multivariate sense the material possesses slightly different characteristics depending on the location it was produced (covariance structure changes with location). Some batches of that raw material was used for a specific product. The material properties after micronization are projected on principal components. It can be observed that after micronization, the batches from a specific supplier location (circled by the small ellipse) project on a different location from the rest. The filling performance of the material originating from this location is also different from the rest of the material. *The raw material variability propagates to quality if the process remains fixed.* A note here that although the control ellipses (large ellipses) shown are set by default in the vendor software, they are not interpretable when there is clustering; the assumptions for the calculation of these ellipses are for process monitoring and not for process exploration where there is intentional variation such as that introduced by design of experiments.

The concept of the design space can be easily understood with the example below, depicted in Figures 44.10 and 44.11. In the figures, we have a process where the raw material is

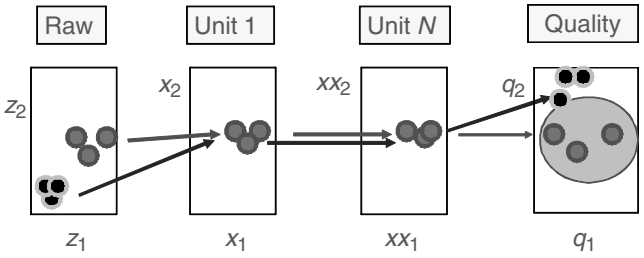

FIGURE 44.10 By maintaining fixed process conditions, we propagate raw material variability to quality.

described by two attributes z_1 and z_2, quality is described by q_1 and q_2, and unit operations described by process parameters x_1 and x_2 for unit 1 and xx_1 and xx_2 for unit N. Two attributes, two quality, and two process parameters per unit are used for illustration purposes, but this does not affect generalization of the following discussion. Each circle represents the values of these parameters for one batch. Figure 44.10 shows what happens when a fixed process is considered, depicted by the Grey circles. Suppose that we run the traditional three batches at a selected range of z_1–z_2 and selected range of process parameters, and we achieve the target quality (all grey circles fall on a multivariate target). The Dark circles represent raw material from, say, a different manufacturer, with attribute values different from the range initially examined. If we process the Dark material on the fixed process conditions (e.g., in the range of the grey circle values), chances are that the final quality will differ from that produced by the grey raw material. Figure 44.11 illustrates that if we carefully choose to operate at appropriate different process conditions for each different material then we can have quality on target. In other words, there is a multidimensional combination of raw material and process parameters that assures quality.

These appropriate process conditions (depicted by the paths that relate raw material and process parameters with quality) are the solutions to the equations of the model that relates raw material and process conditions to quality, and these solutions are obtained when we solve for the values of process conditions given the values of the raw material properties, such that quality falls in a desirable range.

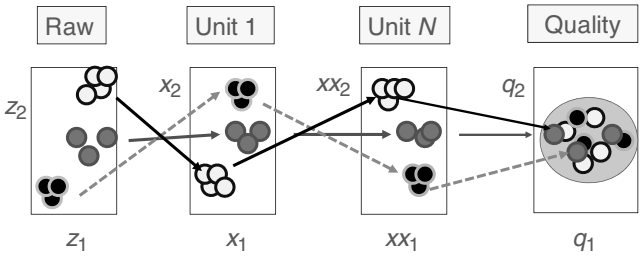

FIGURE 44.11 By taking a feedforward approach where the process conditions are flexible to account for raw material variability, we can maintain quality on target.

Sometimes optimization can be used to introduce constraints, such that the solution takes into account cost, duration of a process, and so on. The model may be theoretical, empirical, or hybrid.

44.9.2 Design Space Modeling

The design space can be established as a model that relates input material and process parameters to quality. The model may be theoretical (based on first principles), or empirical, derived from design of experiments, or a hybrid. Together with the model, one has to specify the range of parameters for which the model has been verified. The model may cover one unit operation or a series of unit operations.

The model will express quality as a function of the raw material attributes and process parameters

$$\text{Quality} = f(\text{raw material},\ \text{process parameters})$$

or, more specifically, as

$$[q_1, q_2, \ldots, q_N] = f(z_1,\ z_2, \ldots, z_K,\ x_1, x_2, \ldots, x_M, \ldots, xx_1, xx_2, \ldots, xx_P)$$

and then solve for the combination of process parameters that will result in a desired $q1$, $q2$, q_N given the values of z_1, z_2, z_N. The function may be linear or nonlinear, and more than one models will in general be required to describe the behavior of a multiunit plant if we wish to be able to predict intermediate quality as well (i.e., granule properties). Multivariate projection methods can be used for empirical modeling. That is, the design space is a collection of models that relate (1) the final quality to all previous units, raw material, and intermediate quality; (2) intermediate quality to previous unit operations and raw material. The design space consists of models (relationships/paths) plus the range of parameters for which the models have been verified. Random combinations in the range of parameters will not work in general (i.e., in Figure 44.11, dark raw material operated in grey conditions may not result in acceptable quality). Therefore, the range without the paths or multidimensional combinations (the model) cannot describe the design space, unless it is the range selected with traditional approaches that describe a fixed process. This range (in this case, grey circle path) may however be only one of the acceptable solutions and therefore restricts our flexibility in dealing with a wide variety of raw materials and/or dealing with disturbances that affect the process.

The above function uses more than two attributes and process parameters in the multidimensional relationships to reflect a general case. Multivariate projection methods or latent variables are proven very useful to describe and solve for these relationships where many variables are present.

The effect of raw material on the quality as it propagates through different unit operations is shown for a tableting process in Figure 44.12. When the raw material properties have certain characteristics (marked with a small ellipse), the material projects on a different area. The properties of granules produced from raw material with such characteristics (black) are different from the rest, and the final quality also shows differences. The difference in the quality can be theoretically explained based on the physical phenomena that govern the whole process. The idea of the design space is to express these phenomena by a model.

Recognizing the continuum in drug production that spans from the drug substance to the drug product will help create a more versatile and robust design space. The final product that delivers the active pharmaceutical ingredient to a patient is indeed the result of a multidimensional combination of raw material attributes and process parameters that span several unit operations including those of the drug substance production (such as reaction and crystallization), the ones from drug product production (such as granulation and compression), and also packaging. Each one of them has an impact not

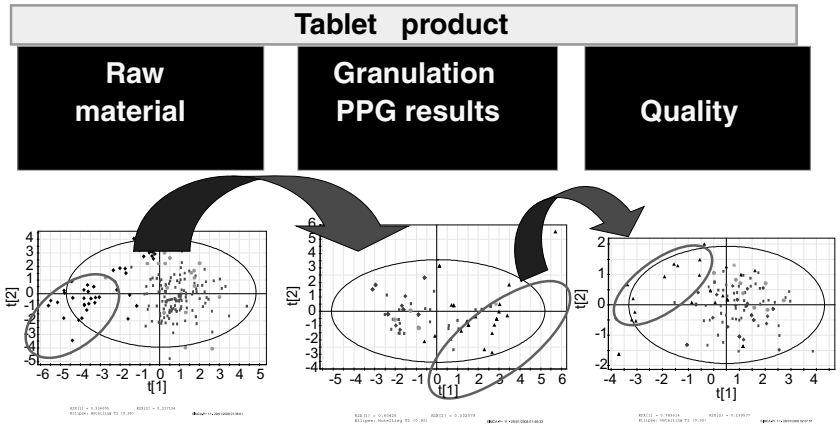

FIGURE 44.12 Projection space representation for a tablet product. Batches produced from raw material with similar characteristics have similar final quality.

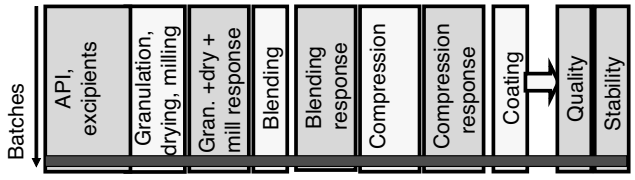

FIGURE 44.13 Both quality and stability profiles can be modeled as a function of input material, process variables, and intermediate attributes.

only on one or more final quality characteristics, but also on stability.

Incorporating stability into the quality by design framework has been discussed; it involves including stability time profiles in the model for the design space [45].

Therefore, if the design space is addressed in a holistic form, then the quality of the final product, including its stability, should be expressed as a function of raw material characteristics and process parameters. Also, by treating the design space in a holistic way, it would provide the manufacturer with the most cost-efficient operation and guarantee high yield and low operating costs because problems at later unit operations will be anticipated and corrected in earlier operations. In other words, the control strategy will be part of the design, such that it can be implemented in the most cost-effective way.

A model that describes the design space for the entire tableting process can be derived by relating quality to both the raw material properties and the process parameters of the unit operations (Figure 44.13). One row in the database depicted in Figure 44.13 would include the process conditions and quality experienced by the product as it is processed through the units. The empirical models derived are causal and based on carefully designed experiments (DOE). Some DOEs will also be necessary to estimate parameters even if mechanistic models are used.

The level of detail in the models varies depending on the objective of the model and the depth of process understanding one wishes to achieve. For example, the variable trajectories of a granulation may be described by carefully selected summary data or by the full variable trajectories aligned against time or another indicator variable.

44.9.3 Control Strategy

Based on the process understanding gained from the design space modeling, the control strategy can be derived to assure final quality. There are several ways of controlling a process, as discussed in Section 44.7. If we decide to keep the process fixed, we may apply a control strategy for the incoming material to reduce raw material variability (see Section 44.7.4).

If we wish to apply the principles depicted in Figure 44.11, then feedforward control should be applied (Section 44.7.1).

Figure 44.14 depicts action in feedforward control that would apply in the case of Figure 44.12. When a different raw material enters the process, we have the choice to adjust granulation conditions in a feedforward manner. However, we also have the choice to adjust compression, that is, to apply the feedforward action at a later stage. When a deviation in the granules is detected that may result in quality different from that typically observed if the compression operates at certain conditions, we may bring the quality on target by altering the compression settings. The choice of the process conditions at unit operation at which we will perform the action will be dictated by a model that takes into account the value of the properties of the input to that unit and calculates process conditions such that quality is on target. When the model is empirical, multivariate analysis can be used.

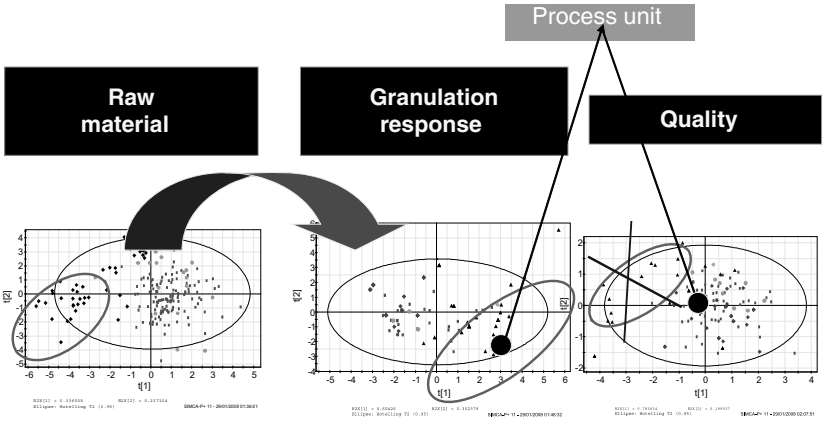

FIGURE 44.14 Feedforward control. When a deviation in the granules is detected that may result in quality different from that typically observed if the next process operates at given conditions, we may bring the quality on target by altering the process conditions.

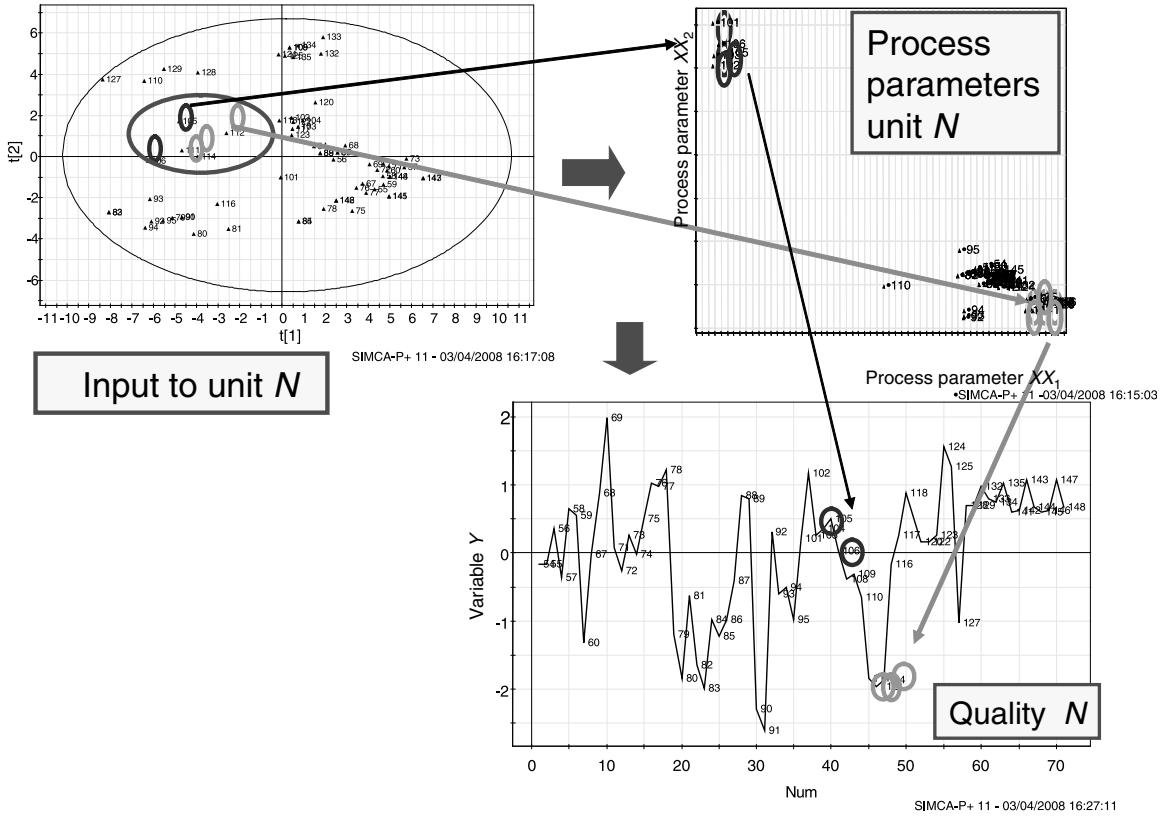

FIGURE 44.15 Control strategy using projection space.

Figure 44.15 illustrates such a case. The example here illustrates a feedforward control scheme for unit N based on input information on the "state of the intermediate product" from unit $N-1$. The settings are calculated and adjusted such that the target value for quality Y is met. A multivariate model was built (from batch data) to relate product quality to the process parameters of unit N and the state of the intermediate product from unit $N-1$, (i.e input to unit N). From this model, a quantitative understanding was developed showing how process parameters in N and the state of the intermediate product from $N-1$ interact to affect quality. Using multivariate analysis assures that the multivariate nature of quality is respected. In this case, the input to unit N is such that, the five batches that project in an area within the small circle (two dark batches and three grey) have the same state of intermediate product—meaning that up to that time the five batches experienced same raw material and processing conditions. The grey batches when processed with typical operating conditions in unit N, marked grey, resulted in quality below average. By taking a feedforward action and processing the dark batches with different operating conditions, in unit N, the quality improves with values above average.

For real-time monitoring and control of an individual unit operation, for example, batch granulation, the principles described in batch process monitoring (Section 44.5.4) and process control by manipulating multivariate trajectories (Section 44.7.3) apply.

44.9.4 Design Space Management

It is accepted that the design space will evolve after the initial submission, and therefore design space management is very important in the product life cycle. There are several issues to consider with design space management, beyond the obvious ones (i.e., beyond managing the design space at the current site to address issues not considered because of limited data available by the initial submission). These include situations where a larger scale is considered at the same site, as well as when there is site transfer. Other issues would be situations where there are different suppliers of API or excipients and when the raw material characteristics are altered slightly within the same supplier. Production changes, such as opportunities to use soft sensors instead of real-time analyzers or to expand the current process analytical technology capabilities, should also be considered. Solutions to such problems can be addressed under the framework of design space management.

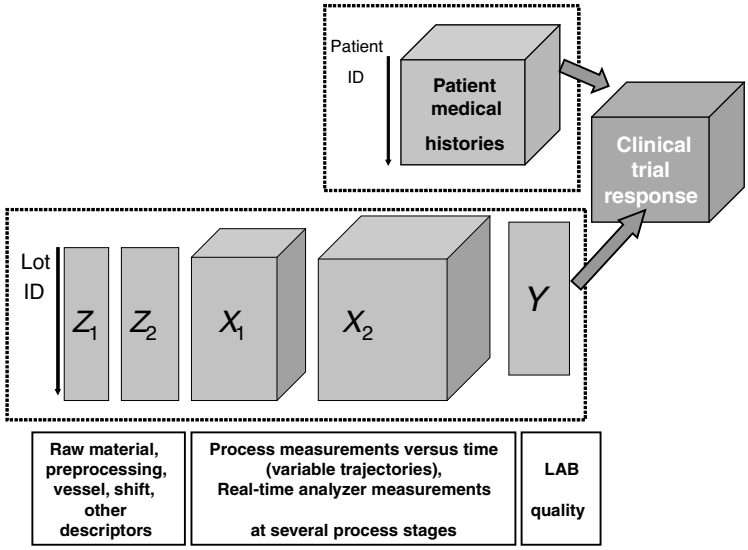

FIGURE 44.16 Examples of complex data structures emerging in industry that can be mined for a wealth of information.

44.10 FUTURE DIRECTIONS

44.10.1 Integration of Clinical Trials

As more complex structures of data are being generated, the multivariate analysis offers great opportunities for information integration and analysis. Both manufacturing data and patient histories can be integrated and then the clinical trial responses incorporated into design space.

Figure 44.16 shows an example of the possibilities that can be explored. Quality in product **Y** can be related to past information of raw materials, preprocessing and holding times, the type of the vessel used, the operator that run the process, and other recipe information, as well as process measurement trajectories and analyzer information. The quality **Y** (and details of manufacturing), as well as the patient medical histories and clinical responses, can be used to establish a better understanding of the design space.

44.10.2 Quality by Design in Analytical Methods

The methodology described for design space can be applied in analytical methods. Chromatography is not only a laboratory method but also a unit operation in biopharmaceuticals. Process transfer ideas can be also applied in method transfer ideas; in other words, method transfer and site transfer could be treated with similar principles.

MSPC for analytical methods has been reported. Multivariate monitoring of a chromatographic system has been carried out using a check sample containing five analytes to test column performance (Nijhuis et al. [46]). A T^2 chart and an SPE_X chart were used to monitor analyte peak area percent of the five analytes. The results indicated that false alarms that would have occurred with univariate charts were avoided and points out of control due to change in correlation could be detected (impossible with univariate charts).

ACKNOWLEDGMENT

The author would like to acknowledge Gordon Muirhead and Bernadette Doyle, GlaxoSmithKline, for their continuous mentoring and support.

REFERENCES

1. Kourti T. Symposium report on abnormal situation detection and projection methods. *Chemom. Intell. Lab. Syst.* 2005;76: 215–220.
2. Abboud L, Hensley S.Factory Shift: New Prescription For Drug Makers: Update the Plants; After Years of Neglect, Industry Focuses on Manufacturing; FDA Acts as a Catalyst; The Three-Story Blender, Wall Street J., Eastern edition, New York, NY, September 3, 2003, page A.1.
3. Yu H, MacGregor JF, Haarsma G, Bourg W. Digital imaging for on-line monitoring and control of industrial snack food processes. *Ind. Eng. Chem. Res.* 2003;42: 3036–3044.
4. Yu H, MacGregor JF. Multivariate image analysis and regression for prediction of coating content and distribution in the production of snack foods. *Chemom. Intell. Lab. Syst.* 2003;67:125–144.
5. Kourti T. Process analytical technology beyond real-time analyzers: the role of multivariate analysis critical reviews in analytical chemistry. *Crit. Rev. Anal. Chem.* 2006;36: 257–278.

6. Kourti T. Multivariate statistical process control and process control, using latent variables. In: Brown S, Tauler R, Walczak R, editors, *Comprehensive Chemometrics*, Vol. 4, Elsevier, Oxford, 2009, pp. 21–54.
7. Brown S, Tauler R, Walczak R, editors. *Comprehensive Chemometrics*, Vol. 1–4, Elsevier, Oxford, 2009.
8. Miletic I, Boudreau F, Dudzic M, Kotuza G, Ronholm L, Vaculik V, Zhang Y. Experiences in applying data-driven modelling technology to steelmaking processes. *Can. J. Chem. Eng.* 2008;86:937–946.
9. Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis using multivariate projection methods—a tutorial. *Chemom. Intell. Lab. Syst.* 1995;28:3–21.
10. Muirhead GT. Process analytical technology at GlaxoSmithKline. Presented at the *19th IFPAC, International Forum Process Analytical Technology*, Arlington, VA, January 10–13, 2005.
11. Graffner C. PAT—European regulatory perspective. *J. Process Anal. Technol.* 2005;2(6):8–11.
12. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J. Chemometrics* 1996;10: 31–45.
13. Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. *Int. J. Adapt. Control Signal Process.* 2005;19: 213–246.
14. Kourti T. Process analysis and abnormal situation detection: from theory to practice. *IEEE Control Syst*. 2002;22(5):10–25.
15. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995;37(1):41–59.
16. Nomikos P, MacGregor JF. Monitoring of batch processes using multi-way principal component analysis. *AIChE J*. 1994;40 (8):1361–1375.
17. Nomikos P, MacGregor JF. Multiway partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* 1995;30:97–108.
18. Kourti T. Multivariate dynamic data modelling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemom.* 2003;17:93–109.
19. Kourti T. Abnormal situation detection, three way data and projection methods—robust data archiving and modeling for industrial applications. *Annu. Rev. Control* 2003;27(2):131–138.
20. García-Muñoz S, Kourti T, MacGregor JF, Mateos AG, Murphy G. Troubleshooting of an Industrial Batch Process Using Multivariate Methods. *Ind. Eng. Chem. Res.* 2003;42:3592–3601.
21. Camacho J, Picó J, Ferrer A. Bilinear modelling of batch processes. Part I: theoretical discussion. *J. Chemom.* 2008;22(5):299–308.
22. Camacho J, Picó J, Ferrer A. Bilinear modelling of batch processes. Part II: a comparison of PLS soft-sensors. *J. Chemom.*, 2008;22(10):533–547.
23. Neogi D, Schlags CE. Multivariate statistical analysis of an emulsion batch process. *Ind. Eng. Chem. Res.* 1998;37:3971–3979.
24. Kassidas A, MacGregor JF, and Taylor PA, Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 1998;44:864–875.
25. Taylor PA.Computing and Software Department, McMaster University, Hamilton, Ontario, Canada, personal communication, May 1998.
26. García-Muñoz S, MacGregor JF, Neogi D, Latshaw BE, Mehta S. Optimization of batch operating policies. Part II. Incorporating process constraints and industrial applications. *Ind. Eng. Chem. Res.* 2008;47(12):4202–4208.
27. García-Muñoz S, MacGregor JF, Kourti T. Model predictive monitoring for batch processes with multivariate methods. *Ind. Eng. Chem. Res.* 2004;43:5929–5941.
28. Kosanovich KA, Piovoso MJ, Dahl KS. Multi-way PCA applied to an industrial batch process. *Ind. Chem. Res.* 1995;35:138–146.
29. Lennox B, Montague GA, Hiden HG, Kornfeld G, Goulding PR. Process monitoring of an industrial fed-batch fermentation. *Biotechnol. Bioeng.* 2001;74(2):125–135.
30. Kourti T, Nomikos P, MacGregor JF. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process Control* 1995;5:277–284.
31. Nomikos P. Detection and diagnosis of abnormal batch operations based on multiway principal component analysis. *ISA Trans.* 1996;35:259–267.
32. Duchesne C, MacGregor JF. Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemom. Intell. Lab. Syst.* 2000;51:125–137.
33. Westerhuis J, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* 1998;12: 301–321.
34. Qin JS, Valle S, Piovoso MJ. On unifying multiblock analysis with application to decentralized process monitoring. *J. Chemom.* 2001;15:715–742.
35. Höskuldsson A. Multi-block and path modelling procedures. *J. Chemometrics* 2008;22(11–12): 571–579.
36. Westerhuis JA, Coenegracht PMJ, Coenraad FL. Multivariate modelling of the tablet manufacturing process with wet granulation for tablet optimization and in-process control. *Int. J. Pharm.* 1997;156:109–117.
37. Findlay P, Morris K, Kildsig D. PAT in fluid bed granulation. Presented at *AIChE*, San Francisco, CA, 2003.
38. Marjanovic O, Lennox B, Sandoz D, Smith K, Crofts M. Real-time monitoring of an industrial batch process. Presented at CPC7: *Chemical Process Control*, Lake Louise, Alberta, Canada, January 8–13, 2006.
39. Zhang H, Lennox B. Integrated condition monitoring and control of fed-batch fermentation processes. *J. Process Control* 2004;14:41–50.
40. Flores-Cerrillo J, MacGregor JF. Control of batch product quality by trajectory manipulation using latent variable models. *J. Process Control* 2004;14:539–553.
41. Dushesne C, MacGregor JF. Establishing Multivariate Specification Regions for incoming materials. *J. Qual. Technol.* 2004;36:78–94.
42. Box GEP, Hunter WG, Hunter JS, Statistics for Experimenters. An introduction to Design, Data Analysis and Model Building.

John Wiley & Sons, New York. Wiley Series in Probability and Mathematical Statistics, 1978.

43. García-Muñoz S, MacGregor JF, Kourti T. Product transfer between sites using Joint Y_PLS. *Chemom. Intell. Lab. Syst.* 2005;79:101–114.
44. International Conference on Harmonization, ICH Draft Step 4, Q8(R2) Pharmaceutical development August 2009.
45. Kourti T, Gonzalez S, Balaguer P, Frances C, Paris G. Stability in the QbD framework. Presented in IFPAC 2010, Baltimore, MD, February 2–4, 2010.
46. Nijhuis A, de Jong S, Vandeginste BGM. Multivariate statistical process control in chromatography. *Chemom. Intell. Lab. Syst.* 1997;38:51–62.