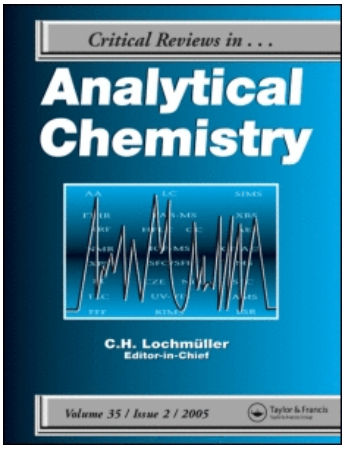

Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Process Analytical Technology Beyond Real-Time Analyzers: The Role of Multivariate Analysis

Theodora Kourti[a]
[a] Chemical Engineering Department, McMaster University, Ontario, Canada

To cite this Article Kourti, Theodora(2006) 'Process Analytical Technology Beyond Real-Time Analyzers: The Role of Multivariate Analysis', Critical Reviews in Analytical Chemistry, 36: 3, 257 — 278
To link to this Article: DOI: 10.1080/10408340600969957
URL: http://dx.doi.org/10.1080/10408340600969957

Critical Reviews in Analytical Chemistry, 36:257–278, 2006

ISSN: 1040-8347 print / 1547-6510 online
DOI: 10.1080/10408340600969957

Process Analytical Technology Beyond Real-Time Analyzers: The Role of Multivariate Analysis

Theodora Kourti
Chemical Engineering Department, McMaster University, Ontario, Canada

Process analytical chemistry was recognized by Callis et al. (*Analytical Chemistry*, 59 (1987): 624A–635A) (1) as a field that extends well beyond real time measurements of process parameters. Process Analytical Technology is taking central stage with the 2004 guidance from the Food and Drug Administration, with a mandate much wider than real time measurements. The pharmaceutical industry is entering a new era. Chemometrics has played an integral part for the real time development of process analytical measurements (multivariate calibration) and it is ready to face the challenge of Process Analytical Technology in this wider definition. The scope of this paper is to demonstrate that multivariate, data based statistical methods, can play a critical role in process understanding, multivariate statistical process control, abnormal situation detection, fault diagnosis, process control and process scale-up, as linked to process analytical technology.

Keywords process analytical technology, multivariate statistical process control, scale-up, latent variables, process understanding, multivariate analysis

INTRODUCTION

Historically, the term process analytical technology (PAT), as stated by Workman et al. (2), "has continued to evolve as a more appropriate term than process analytical chemistry (PAC) to describe the field of process analysis, as measurement technologies are expanding to include many physical characterization tools. This term has existed since the turn of the century (ca. 1911) but is only now found in common usage."

Recently PAT has been assigned the following definition (3, 4) by the Food and Drug Administration (FDA): "systems for the analysis and control of manufacturing processes based on timely measurements during processes of critical quality parameters and performance attributes of raw and in-process materials and processes to assure acceptable end product quality at the completion of the process." Under the recently reissued current Good Manufacturing Practice (cGMP) guidelines (4) FDA expects that PAT will be essential for process understanding and will result in improved process control strategies for high-quality, cost-effective pharmaceutical products.

Kourti (5) argued that "the approach described as PAT by FDA is an approach that has been taken by several other industries (petrochemical, polymer, chemical) several years ago. Real-time measurements were collected during processes on process variables, other parameters, as well as on quality by real time analyzers. Process control techniques were developed. Attempts were made to understand the fundamental mechanisms of processes and to build models. But it was not until 15 years ago, that the industry witnessed yet another revolution and managed to achieve (and exceed) the goals described by FDA's PAT definition. What made the difference was the ability to perform multivariate statistical process control (MSPC) and monitor the wellness of the process and product, by looking simultaneously at hundred of variables collected in real time."

From a historical perspective, it is worthwhile revisiting the statements made, almost 20 years ago, by Callis et al. (1): "The goal of process analytical chemistry is to supply quantitative and qualitative information about a chemical process. Such information can be used not only to *monitor* and *control* a process, but also to *optimize* its efficient use of energy, time, and raw materials. In addition, it is possible to simultaneously minimize plant effluent release and improve product quality and consistency." The authors then continue discussing the potential impact of process analytical chemistry on the traditional production of basic chemicals: "The emphasis is shifting from adding capacity to making existing plants more efficient to meet the demands of increasing international competition. This competition involves not only pricing; in many cases the issues of product quality and consistency are equally important. Greater emphasis has been placed on lowering the "quality cost" of production—the loss

Address correspondence to Theodora Kourti, Chemical Engineering Department, McMaster University, 1280 Main West, Hamilton, Ontario, L8S 4L7, Canada. E-mail: kourtit@mcmaster.ca

suffered when a batch of product fails to meet specifications and has to be reprocessed, sold at reduced cost, or discarded." And they add: "However, one must realize that the practice of process analytical chemistry involves much more than improved chemical sensing. The issues of sampling extraction of information from the sensor data, integration of the information into process control, and the sociological difficulties of gaining process engineers' and plant operators' confidence in the new measurement tools must all he given equal consideration. Clearly, process analytical chemistry should be considered a worthy sub discipline of analytical chemistry, and it requires an interdisciplinary systems approach so that progress can be made."

Notice that if one replaces the words "production of basic chemicals" with "production of pharmaceutical products" Callis et al. (1) practically gave the same definition for the role of PAT as now does the FDA (3) and cited the same motives for its need.

Although this paper covers PAT beyond real-time analyzers, for completeness of the subject, the reader is also alerted to the recent reviews on both Chemometrics and Process Analytical Chemistry. These reviews deal mainly with new technology for real time measurements and chemometric methods related to calibration modeling. Process analytical chemistry, has been reviewed in a series of six articles (2, 6–10), published since 1993 every two years in the journal of Analytical Chemistry; the last one (2) covers the period from March 2003 through March 2005. The authors cover analytical technologies applicable toward process, remote, real-time, and high-throughput measurements. They review key developments for an extensive array of measurement techniques as well as those traditionally associated with process analytical chemistry. They also address advances in nanotechnology, microelectromechanical systems (MEMS), and microanalytical systems. Research developments in process analytical chemistry over that period of two years, are also included. Recent reviews on chemometrics as calibration methods for process analytical instrumentation can be found in (11–15). A recent book (16), written from the perspective of the spectroscopist required to implant PAT tools in a process environment, covers key spectroscopic tools, their applications in the pharmaceutical and chemical industries and basic chemometrics. Finally in depth discussions for specific tools also exist. NIR for example, that has been one of the fastest growing process analytical technologies, is discussed by Siesler (17) who explains that the availability of efficient multivariate calibration methods, light fiber optics coupled with specific probes and miniaturization has made NIR spectroscopy an excellent tool in industry for assessing quality and for process control. Calibration transfer is of course a hot topic for on-line instrumentation. An overview of the different methods used for calibration transfer and a critical assessment of their validity and applicability, with focus on methods for NIR spectra can be found in Feudale et al. (18).

This paper provides a critical discussion on the potential of data based multivariate statistical methods to meet the PAT challenge and on the tools required to address specific areas within the PAT frame. Such areas include process understanding, multivariate statistical process control, abnormal situation detection, fault diagnosis, monitoring analyzer reliability, soft sensors, inferring final product quality from process conditions during production, establishing an overall "process signature," process control and process scale up.

Due to space restrictions, the presentation of detailed equations and algorithms is limited, and references are given instead for tutorials and other essential papers. Wherever there are many references available for a subject, referencing is restricted to the most recent application work and to the work that covers the corresponding theoretical development.

CRITICAL CONCEPTS

Multivariate Nature of Quality

The multivariate nature of quality should become the foundation for approaching the whole concept of PAT. 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 1 is a classic illustration of the problem with using separate control charts for two quality variables (y_1, y_2). It is repeated here because it will form the basis for several discussions. 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 (19). 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-α)% joint confidence limit of the distribution (i.e., when the process is in control, α% of the points will fall outside the ellipse).

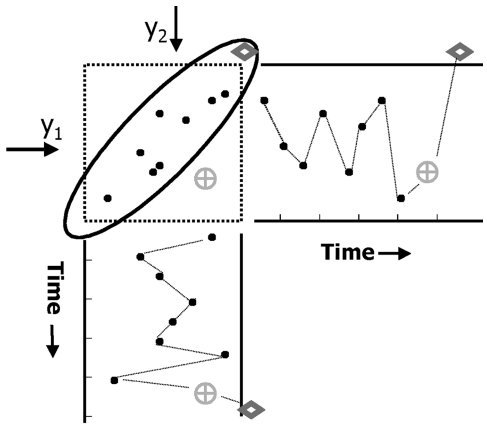

FIG. 1. Plotting variables y_1 and y_2 against each other, reveals abnormalities that are not seen by plotting the variables individually on univariate control charts.

The point indicated by the ⊗ 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 ⊗; 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 ⊗), but also rejecting a good product as bad (point ◇). The problem worsens as the number of variables increases. It is clear that an efficient fault detection scheme should look at the variables together. Multivariate charts (as for example Hotelling's T^2) are required in order to test quality (20). Multivariate charts of lab data entries, offer the additional advantage of quick detection of outliers (e.g., erroneous data entries due to human error).

Recognizing the multivariate nature of quality, makes it necessary to adopt the following policies:

- Requirement for Multivariate Control charts for Raw Material Evaluation.
- Requirement for Multivariate Control charts for Quality Property testing in all intermediate steps of the process.
- Requirement for Multivariate Control charts for Testing Final quality; product release should be based on this multivariate test.
- Requirement for Multivariate Charts for product transfer and scale up (for all raw materials, intermediate qualities and final qualities). This is a minimum requirement scale up. Later, we discuss the requirements on the multivariate space of the process variable trajectories as well.

Accordingly the definition for Critical Quality attributes should be corrected as shown in Table 1.

What Defines Final Product Quality—Is Meeting Product Specifications Sufficient?

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:

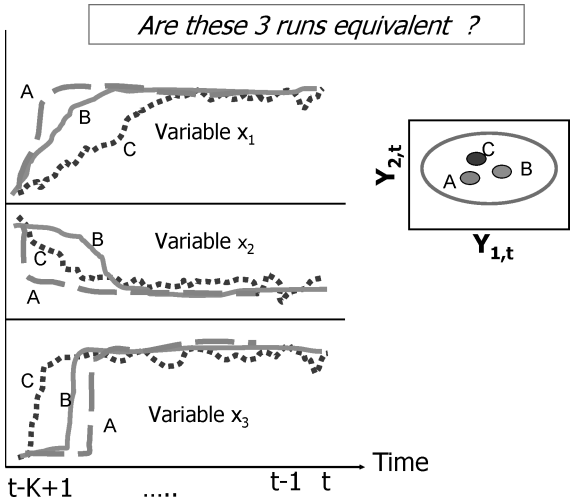

FIG. 2. The trajectories of 3 process variables for 3 different batch runs (A, B, C) follow different paths. There is a very high probabilty that the overall quality (and downstream processability) for lots A, B and C is different, although two "end properties", $y_{1,t}$ and $y_{2,t}$, are on target in the multivariate control charts.

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

Consider the example of Figure 2, where the trajectories of each one of three process variables are plotted for three different batch runs, A, B, C. The final product at 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, 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? As discussed later, there is a very high probability that they are not.

The fact is, that very frequently, the few properties measured on a product are insufficient to define entirely the product quality.

TABLE 1
The definition of Critical Quality Attributes should be modified to take into account the multivariate nature of quality.

Current definition	Correct definition
Critical Quality Attribute (CQA)—A measurable property of an intermediate or final product that is considered critical for establishing the intended purity, efficacy, and safety of the product. That is, the property must be within a predetermined range to ensure final product quality.	Critical Quality Attributes. The combination of measurable properties of an intermediate or final product that are considered critical for establishing the intended purity, efficacy, and safety of the product. That is, the value of the multivariate statistic calculated for these properties must be within a predetermined limit of a multivariate chart to ensure final 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. However, process data (temperatures, pressures, etc.) collected during the polymer production may contain valuable information about events with special causes that may affect the product chemical structure and thus its performance in different applications.

Another example that further corroborates this argument has been reported from the Pharmaceutical Industry (21): "For a given pharmaceutical unit process, the determination of "end point," and the ability to stop the process at this event, are critical. Conventional process control of, for example, 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 on-line 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 (22).

Dushesne et al. (23) studied this topic extensively for grade transitions. Grade-to-grade transitions are non-linear processes and they can be formulated and modelled by multivariate methods the same way as batch processes. The authors demonstrated that the same measured quality properties can be achieved by taking different process paths. However, 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 was not the case, although the measured product properties were on target, the properties that determined the processability of the product were not within acceptable limits.

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, etc., combined with real time measurements from analytical technology (spectroscopy, ultrasound, etc) will lead to on-line *process* monitoring and make multivariate statistical process control (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 is pointed out later that sometimes real-time process measurements may eliminate the need of some real-time analysers. Information about the process may also include the vessels used for a specific run, the operators that were on shift, suppliers of raw material, etc. 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 as illustrated later.

Process data can be utilised together with appropriate chemometric 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
- Decide on mid course 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, once we are in production, these relations may be influenced by other factors and change. An example of the effects on these relations when the system works with all the controllers in place is given in (24); 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 in SPC. It should be emphasized here that the parameters that appear to be critical in the design of experiments (DOE), are not necessarily critical 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 non-causal; 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. That 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.

Multivariate Nature, Structure and Other Characteristics of Process Data

Databases containing measurements collected during production may become very large in size. The data are non-causal 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 paper in expressing data is that of Table 2. Other formats which may appear in specific sections only, will be defined in their corresponding sections.

Process Analysis and Process Understanding

In their paper Callis et al. (1) described the "five eras" of process analytical chemistry as "off-line, at-line, on-line, in-line, and noninvasive." In the paper the authors made it clear that this definition was referring to the way measurements were collected on a property of interest. In their words, "the first two eras, require the manual removal of the sample and transport to the measuring instrument"; In the on-line era of process analytical chemistry "an automated sampling system is used to extract the sample, condition it, and present it to an analytical instrument for measurement; It is possible to subdivide the online era into two categories: intermittent methods that require injection of a portion of the sample stream into the instrument and continuous methods that permit the sample to flow continuously through the instrument." "In-line chemical analysis is done in situ, directly inside the process line, using a probe that is chemically sensitive. In its ideal form, the device might resemble a typical industrial temperature probe"; "The final era of process analytical chemistry, the noninvasive era, represents the ultimate in desirability. Because the probe does not physically contact the sample, the sampling problem is greatly alleviated. This era obviously has a great deal in common with remote sensing and nondestructive evaluation. Near-IR spectroscopy from 700–1100 nm has much to offer in this regard."

Unfortunately that definition has been misinterpreted through the years and as a result we read in the literature that Callis defined "five areas" or "five categories of process analysis" and that the "last there categories form the basis of what is now commonly described Process Analysis."

Process analysis should not be used as an equivalent term to process analytical chemistry. It has a much wider definition than this in the scientific and engineering community. Collecting real-time measurements on a specific property may or may not reflect what is happening in the process or, the state of the process, as discussed earlier; therefore collecting a real-time measurement is not necessarily 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

TABLE 2

Matrix symbol	Dimensions	Explanation
X	*(n × 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.)
Y	*(n × 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 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.
Z	*(n × r)* *n observations in time, or n batches; r other variable measurements*	Raw material and other pre-processing information

to develop ideas to improve it. Improvement could translate to better quality, lower cost, more efficient energy consumption, less pollutants to environment, 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 the process can be described by first principles, that is by a theoretical or mechanistic model. This entire paper is based on the philosophy (and also demonstrates) that one can gain tremendous insight in to the process from empirical models derived from data bases; as a matter of fact this is also the basis of chemometrics. These empirical models can lead to fast improvements that in several situations would have been impossible had people been waiting for the development of theoretical, first principles models. It will also be discussed that 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:

- 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 the process data of the last month, projected on a latent variable space, seem to form two clusters indicating different operation practices and 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 re-sized.)
- Assessing that an operational problem caused the readings of a specific thermocouple to be erroneous and to appear as outliers (The thermocouple was too close to the entrance of cool monomer; erroneous readings fed to the controller alter inappropriately 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 production scale and understanding equipment related operational problems is a tremendous asset to the effort of product quality improvement and sometimes it weighs equally important to understanding the detail mechanisms of the reaction that takes place.

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 (25):

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} \quad [1]$$

$$\mathbf{Y} = \mathbf{TQ^T} + \mathbf{F} \quad [2]$$

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 if 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., non-full 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 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. This is also illustrated in a relatively old publication (26) while Sutter et al. (27) suggested a methodology to assist at the selection of important components to include in the regression. 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 in (25). 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 in the **X** space as well as in the **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 (28–30). The properties of PCA and PLS are discussed briefly below.

Principal Component Analysis. For a sample of mean centered and scaled measurements with n observations on k variables, **X**, the principal components (PC) 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, etc. Up to k PCs are similarly defined. The sample principal component loading vectors $\mathbf{p_i}$ are the eigenvectors of the covariance matrix of **X** (in practice, for mean centered data the covariance matrix is estimated by $(n-1)^{-1}\,\mathbf{X^T X}$). The corresponding eigenvalues give the variance of the PCs (i.e., var $(\mathbf{t_i}) = \lambda_\mathbf{i}$). 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 Eq. [1].

Partial Least Square Analysis. 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 $\mathbf{X^T Y Y^T X}$. 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, $\mathbf{p}_1 = \mathbf{X}\mathbf{t}_1/\mathbf{t}_1^\mathbf{T}\mathbf{t}_1$; 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 $\mathbf{X}_2 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^\mathbf{T}$. **Q** are the loadings in the **Y** space. In NIPALS, $\mathbf{q}_1$ is obtained by regressing $\mathbf{t}_1$ on **Y,** then **Y** is deflated $\mathbf{Y}_2 = \mathbf{Y} - \mathbf{t}_1\mathbf{q}_1^\mathbf{T}$. 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 $\mathbf{X}_2^\mathbf{T}\mathbf{Y}_2\mathbf{Y}_2^\mathbf{T}\mathbf{X}_2$, and so on. The new latent vectors or scores $(\mathbf{t}_1, \mathbf{t}_2, \ldots)$ and the weight vectors $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ are orthogonal. The final models for **X** and **Y** are given by Eqs. [1] and [2].

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 $(t_1, t_2, \ldots, t_A)$ 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.

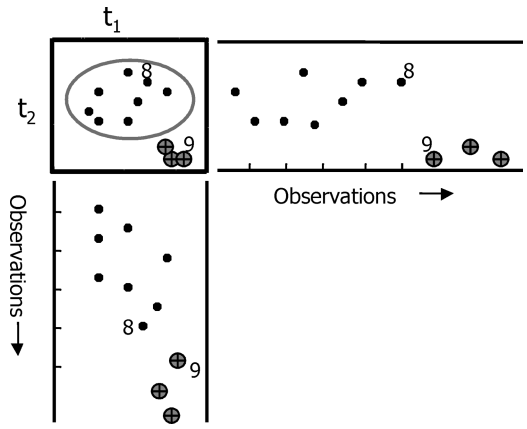

FIG. 3. By plotting scores against each other the process behaviour can be observed at a lower dimensional space.

Outlier and cluster detection becomes also easy, both for the process and 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 vs t_2 plot is the summary of measurements on k variables. (Notice that in Figure 1, when we plot raw variables y_1 and y_2, each point is the summary of the two variables only). Figure 3 illustrates two clusters of points observed on a t_1 vs t_2 plot. The use of contribution plots may help to investigate which variables have contributed to the move from point 8 to 9.

The contribution of variable j to the move of the score values between two observations (say, 8 and 9) for component q is calculated as

$$p_{jq} \times (x_{j,9} - x_{j,8}) \text{ for PCA} \quad \text{and}$$
$$w_{jq} \times (x_{j,9} - x_{j,8}) \text{ for PLS,} \qquad [3]$$

where p_{jq} is the loading of variable j on component q and w_{jq} is the weight of variable j on component q (31).

The use of latent variable methods to explore and analyze historical data is illustrated by a troubleshooting problem on an industrial continuous recovery process with 447 process variables (31). Contribution plots pointed to the source of the problem that was responsible for a drop in recovery for 3 months of operation.

Multivariate Statistical Process Control (MSPC)

From routine operation we can establish acceptable limits of good behavior. On t_1 vs 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 3 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 1) this is not the case

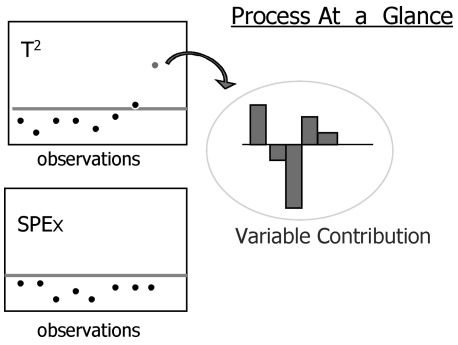

FIG. 4. The Hotelling's T^2 chart and the SPE chart summarize the process behaviour.

when it is calculated for the principal components which are orthogonal.

It would have become cumbersome to have to plot all combinations of principal components (even if we had 4 components we would need 6 charts). Fortunately a statistic (Hotelling's T^2) can be now calculated and the overall performance of the main events of the system can be monitored with a single chart. One such chart is the one shown at the upper left corner of Figure 4. The solid line corresponds to acceptable performance. For the case of two components, the solid line corresponds to the perimeter of the ellipse of Figure 3. For 3 components it would correspond to the surface of an ellipsoid, for 4 components the surface of a hyper ellipsoid.

The Hotelling's T^2 for scores is calculated as:

$$T_A^2 = \sum_{i=1}^{A} \frac{t_i^2}{\lambda_i} = \sum_{i=1}^{A} \frac{t_i^2}{s_{t_i}^2} \qquad [4]$$

where s_{ti}^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 previously 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 4, 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 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:

$$\text{SPE}_X = \sum_{i=1}^{k} (x_{new,i} - \hat{x}_{new,i})^2 \qquad [5]$$

where $\hat{\mathbf{x}}_{new}$ 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 $\mathbf{E}$ in Eq. [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 refence model. The calculation of the limits for the charts is discussed in (20).

These two charts (T^2 and SPE) are two complementary indices; together they can give a picture of the state 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 $(x_{new,j} - \hat{x}_{new,j})^2$. Variables with high contributions are investigated.

Contributions to Hotelling's T^2. Contributions to an out of limits value in the Hotelling's T^2 chart are obtained as follows: a bar plot of the normalised scores $(t_i/s_{ti})^2$ is plotted and scores with high normalised 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:

$$c_j = p_{q,j}(x_j - \bar{x}_j) \text{ for PCA} \quad \text{and}$$
$$c_j = w_{q,j}(x_j - \bar{x}_j) \text{ for PLS} \qquad [6]$$

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 for meancentered data). Variables on this plot that appear to have the largest contributions to it, but also the same sign as the score should be investigated (contributions of the opposite sign, will only make the score smaller). When there are K scores with high values, an "overall average contribution" per variable is calculated, over all the K scores (32).

Utilizing contribution plots an abnormal situation can be detected in real time, the source of the problem can be diagnosed and corrective action may be 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 where the reactor temperature in an exothermic reaction had to be maintained at 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; 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 the factors influencing the process during production and help in an "ongoing process understanding" philosophy.

MSPC for Raw Material and Final Quality Assessment. Multivariate charts can be constructed to assess the consistency of the multivariate quality of raw materials, **Z**, as well as 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.

Reference Data Set for SPC Modeling. When dealing with empirical modeling, the data set on which the model will be based must be chosen carefully to satisfy the needs of the intended application.

For inferential modeling and response surface modeling, one needs data from designed experiments over a prescribed range of **X** and **Y** variables. Typically, a wide range of process conditions is considered in this case so as to choose the optimal operating region. When the model is to be used for SPC, only a specific operating region is tackled.

Historical process and/or product data corresponding to this operating region should be used. The objectives are to model the good process behavior in this operating region and to test for any future deviations from this model. All the data should correspond to in-control operation, and faults or disturbances are excluded from the model. If the preliminary analysis of the historical data in this region indicates clusters containing only a few points or shows individual outliers, these data should not be discarded, but investigated. True outliers (measurement errors) should be discarded if they can be identified. If the small clusters reflect some real unusual event that still produces acceptable product, then the data could be included in the model only if more data points in this region can be collected to establish a robust model. Otherwise these data should be left out during modeling, then tested with the model as if they were new data, to identify what alarm signal they produce (direction of principal components). This information should be stored and used to provide a warning that the particular signal corresponds to this known "rare situation" and not a bad product (31).

Finally, when exploring the process through historical data analysis, all the data should be used initially so that outliers are identified and discarded. The rest of the data should then be used in the projection for the analysis of past behavior.

Observability/Detectability of Faults. When process data are used to derive an empirical model for process monitoring, one should keep in mind that the objective is to detect faults when they occur. Therefore, the model should be tested with known faults to determine the "observability" of these faults. Suppose that the SPE chart and the Hotelling's T^2 at A components will be used to monitor the process. Then, if a set of process data are known to have caused a defective product, when these data are passed through the model, the SPE chart or the Hotelling's T^2 chart, or both, should signal the problem. If they do not, then the specific problem cannot be seen with this model (i.e., this set of charts). Sometimes this means that the measured process variables do not contain the signature of this fault and more representative variables should be collected. Other times (in the author's experience) it requires that certain process variables be given a higher weight in the model. This happens when a set of variables are in tight control under good production (i.e., show very small variation), and as a result, these variables are buried in the SPE (i.e., do not appear in the first major principal components). They then require a large deviation from normal before being detected. To identify these variables, one needs to build a "troubleshooting" model containing both the good data points and the fault and use contribution plots (between good points and faulty points) to isolate these variables. The model is then developed using only the good data points, but giving higher weight to these variables so that they appear on the first major principal components.

Finally, there are situations where it is more important to detect some faults than others. To detect a specific fault, a separate Hotelling's T^2 chart containing the scores where the specific fault manifests itself can be used in addition to the traditional charts described earlier (31).

MSPC in Industry

Traditionally statistical process control in industry, has been synonymous with monitoring product quality variables or some key process variables in a univariate way. The direct result of this is the large number of control charts that are usually present in a control room and that the operators have to attend to. When there is an abnormality in the plant operation several of these charts alarm in a short period form each other, or simultaneously. This happens simply because process variables are correlated, and an abnormal event may affect more than one variable at the same time. When such a situation occurs, it is difficult for the operator to isolate and determine the source of the problem, which may lie with only one of the many correlated alarming variables, or, as it is most frequently the case, it may simply be a non-measurable variable (impurities, plugged pipe, blockage of a sensor) that causes several other measured variables to go out of control.

The use of projection methods has revolutionalised the idea of statistical process control for multivariate processes. The performance of an entire unit, or even a plant, can be monitored by the operator looking at only a few multivariate control charts, that can be thought of as process performance indices. These charts, which are based on latent variables are simple, easy to understand by the operators and have found quick acceptance in

the control rooms. They improve early fault detection capabilities, because they are able to detect the onset of a fault at the same time as, but in most situations earlier than, the many univariate charts. More importantly however, they detect problems that manifest themselves as changes in the covariance structure of the process variables, which univariate charts will miss if the variables remain within their expected univariate operation limits. The methodology based on latent variables also provides diagnostic tools that help the operators determine quickly and efficiently the source of the problem.

The methods and their potential and limitations for improving operations are discussed in Kourti (31, 32). Examples from state of the art major industrial applications currently running online are presented to illustrate the tremendous potential of these methods. In this context, an industrial application for abnormal situation detection is defined as "state of the art," if it has been operational several years after it was commissioned, has generated large savings (sometimes on the order of millions of dollars per year), has been operating safely and/or has improved safety conditions in the plant, and is accepted enthusiastically by the operators. The above articles also contain an extensive literature review on the subject and also include practical considerations for the user (frequency of sampling, utilizing process knowledge) as well as warnings for potential pitfalls, from data acquisition to modeling to on-line application. The industrial practitioners perspective can be found in Miletic et al. (33) and Kourti (34).

MULTIVARIATE ANALYSIS OF BATCH PROCESSES: THREE-WAY MATRICES

Modeling

Batch processes are dynamic processes. The variable trajectories measured over the duration of the batch are non-linear with respect to time and form a multivariate time series with dynamic nature. (Figure 2). The product quality (properties of the product $\mathbf{y}_{it}$) measured at the end of the batch at time t is a function of the process conditions at time t, but also a function of the process conditions prevailing several lags before, and in most cases a function of the conditions prevailing during the entire batch. Take emulsion polymerization for example: the particle size depends strongly on (in fact it is mainly defined by) the nucleation period which takes place at the first few minutes into the reaction. What happens in the rest of the batch will also affect particle size (rate of reaction will affect growth, while agglomeration phenomena and/or secondary nucleation will affect the shape and average size of the particle size distribution). Therefore, the final particle size (quality as measured at the end of the batch/semi-batch process) is a function of the process conditions during the entire batch. Similar is the situation observed in crystallization; crystal size distribution, crystal shape, and polymorphic form are affected by the prevailing conditions for the duration of the batch.

In PCA/PLS inferential modeling, when we deal with dynamic multivariate time series data, in order to relate input $\mathbf{X}$ to output $\mathbf{Y}$, the $\mathbf{X}$ matrix is expanded to include values of the x variables at several lags (35). If the number of lags is equal for all the x variables then the $\mathbf{X}$ matrix corresponding to the data of Figure 2, looks like that of Figure 5, bottom. (In this example $\mathbf{Y}$ is available only at time t; when $\mathbf{Y}$ is available on-line, at several time intervals, the set up can be modified to accommodate that). For a batch process where J process variables are measured at K time intervals, or K aligned observation numbers (A.O.N.), for each one of I batches, the data could also be folded and represented by a three-dimensional data array $\underline{\mathbf{X}}\,(I \times J \times K)$–Figure 5, top.

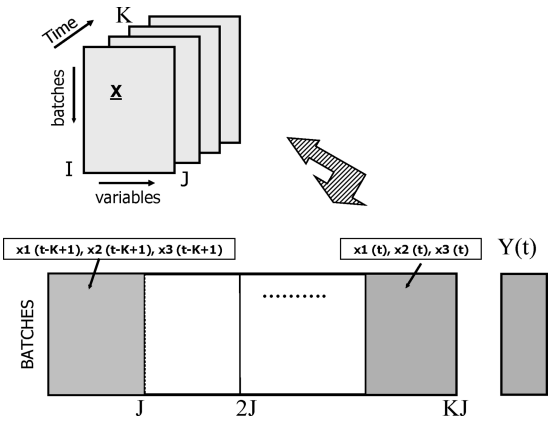

FIG. 5. In dynamic multivariate time series data, in order to relate input $\mathbf{X}$ to output $\mathbf{Y}$, $\mathbf{X}$ is expanded to include values of the x variables at several lags. – Fig 5, bottom. For a batch process where J process variables are measured at K time intervals, for each one of I batches, the data could also be folded and represented by a three-dimensional data array $\underline{\mathbf{X}}(I \times J \times K)$ – Fig 5, top.

Methodology to analyze data bases collected from batch processes is fortunately available. A methodology utilizing multiway principal component analysis and multiway partial least squares was initially presented by Nomikos and MacGregor (36–38) in a series of landmark papers. This methodology applies to operations of finite duration, like batch distillation, batch reactors, batch annealing, mixing additives for a finite time, drying and chromatographic separations.

As discussed in the literature (39, 40), 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 and others are not measured for the full duration of the batch. Furthermore, the frequency of measurements may be different: (i) some variables may be measured more frequently than others (i.e., some every minute and others every 15 minutes); (ii) certain phases in the process may be sampled more frequently to catch important phenomena. The data set in such situations does not form a complete cube, but rather a cube where some columns are missing. Figure 6 shows a simple example of 3 variables in 4 time intervals in a batch process. Variable x_1 is measured at all 4 time intervals, variable x_2 is only

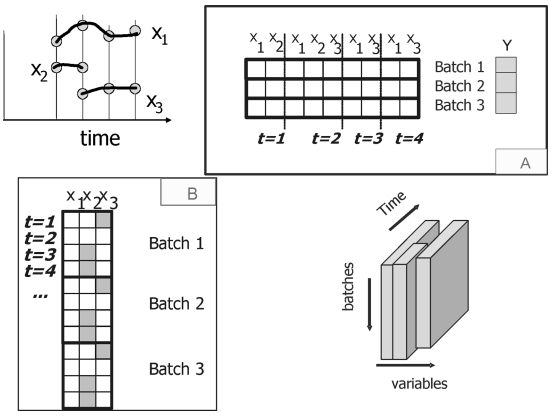

FIG. 6. When variables are not measured for the full duration of the batch, the data set does not form a complete cube. Data can be unfolded "batch wise" as shown in insert A, or "variable wise" as shown in insert B; grey areas in B indicate that data are not available.

sampled at time $t = 1,2$ and variable x_3 is not available at $t = 1$. The lower right corner shows the incomplete cube formed from these data.

The methods that are used to model batch processes should be capable of modeling the structure of this incomplete cube; the one presented by Nomikos and MacGregor (36–38) is capable of modeling such a structure. Furthermore, it is capable of modeling three way structures generated when formulating the control problem of batch processes using latent variables which is discussed later. Finally this unfolding can retain and model the auto and cross correlation of all the variables in the batch for the duration of the batch (39). This is an extremely important issue when using the model for fault detection (enhances fault detection capability) but also when using the model for optimization and control of batch trajectories.

In Figure 6, the method of Nomikos and MacGregor unfolds the data in the way shown in insert A (batch wise). Notice that the variables are arranged next to each other at different times; when a variable is not sampled, the corresponding column simply is not present. Notice that with this unfolding, one row corresponds to one batch and matches the corresponding product quality data (Y). In a different way of unfolding (variable wise) the batches are arranged one below the other. Notice that in this case when the variable is not sampled it is always missing (shown by grey areas in Figure 6). However missing data approaches cannot be used to solve this problem because the same "time–variable" combination is always missing, hence that combination is not observable.

A critical discussion on other batch processes modelling procedures and other issues related to batch process analysis can be found in recent papers (32, 39, 40); mean centering and scaling the incomplete cube is discussed in (32).

Batch Processes of Different Duration

Batch processes often have variable duration; for the same recipe, in the same vessel sometimes we need more time to achieve the same yield. One important stage before modeling batch process data is the alignment or synchronization of the data. With alignment or synchronisation we must achieve the following: (i) Establish common start points at different phases of the run; these start points may coincide with the start of a physical or a chemical phenomenon and (ii) Match the shape of the trajectories of the variables; the trajectories may be expressed as function of time or as a function of another variable. Once the shapes match it is not necessary that the length of the batches match (40). A first attempt for batch data alignment involved the use of the cumulative monomer added to the reactor as an indicator variable (41); 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 (42).

Dynamic time warping, based on speech recognition methods was suggested by Kassidas et al. (43); they also suggested the use of total time as a variable in the **Z** matrix, as extra information to describe the batch. Taylor (44) 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 used later with excellent results (45, 46) and also provided the basis to designing batches with desired duration (47–49); Westerhuis et al. (45) used the cumulative warp as an extra variable to take into account time effects on batch quality, when batches were synchronised by dynamic time warping, while García-Muñoz et al. (46) used the cumulative time when batches were synchronised by the indicator variable approach. A critical discussion of the various synchronization approaches can be found in Kourti (40). Provided that an indicator variable exists (or can be constructed by non-linear transformations from other variables and /or process knowledge), the indicator variable approach is usually chosen as the simplest and most convenient for industrial applications. Other synchronization approaches have also been presented to deal with alignment of chromatographic data (50).

Process Understanding

One of the key issues addressed in FDA's PAT guidance is *process understanding*. Latent variable batch modeling is a superb tool for achieving process understanding. Scores, loadings, and other parameters calculated by the model become extremely useful for this purpose:

- Plots of scores and residuals, calculated from historical data sets of completed batches, will help identify clusters and outliers;

- Contribution plots will help identify reasons for cluster separation
- Plotting the loadings per variable and time and per component would provide valuable information for the covariance structure of the variables for the duration of the batch and will explain physical or chemical phenomena occurring in the batch (39).

A detailed tutorial of how multivariate analysis can be used in troubleshooting and understanding of a batch process with different phases is presented in (46); the paper deals with issues like alignment of different phases, classifying good and bad runs and troubleshooting; utilizing information from raw data using multiblock methods; detecting clusters of good and bad batches; utilizing contribution plots to interrogate the multivariate model and determine combinations of variables and periods of operation that will drive a process away from producing a good product and towards producing defect. Based on this analysis, it was possible to determine raw material combinations and processing conditions that would result in a bad product.

Multiway methods and design of experiments can be used (52) to determine optimal process variable *trajectories* in a batch process in order to obtain a desired quality property.

MSPC in Batch Processes

Multivariate control charts (Hotelling's T^2 and SPE) can be constructed for batch processes in a straight forward manner (36–38). Therefore MSPC is possible. 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).

Industrial Practice. Industrial applications for batch analysis, monitoring and fault diagnosis have been reported (42, 46, 52–56). It should be noted here that several companies choose to use the methodology not for on-line monitoring but as a tool for *real-time release* of the batch product. That means that the batch is not monitored as it evolves; 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 hours, but the product analysis may take 8 hours. That means that they do not have to waste 4 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.

Industrial applications where by monitoring the batch process variables abnormal operating conditions (a pending equipment failure) were detected, were also reported (54). The same methodology that it is used for batch monitoring it is also used for start ups of continuous processes. An excellent application that details the procedure for on-line implementation and fault detection can be found in (57).

MULTIBLOCK ANALYSIS FOR MULTI-UNIT, MULTIPHASE, OPERATIONS

Very frequently in industry one encounters processes that involve several unit operations. For example, as shown in Figure 7, we may have two batch processes in sequence ($\underline{\mathbf{X}}_1, \underline{\mathbf{X}}_2$) followed by a continuous process $\mathbf{X}_3$. Furthermore information may be available on raw materials $\mathbf{Z}_1$, and other preprocessing data $\mathbf{Z}_2$. All of these have an effect on the final quality, $\mathbf{Y}$.

The data from all the units can be incorporated into one model for the entire process; this model will take into account the interactions between units and their relative importance to the final product quality by weighing them differently. This is the approach of multiblock PLS (MB-PLS) (51, 58). MB-PLS is not simply a PLS between each $\mathbf{Z}_i$, $\mathbf{X}_i$ block and Y. The blocks are weighted in such a way that their combination is most predictive of $\mathbf{Y}$. Superscores are generated describing the behavior of the

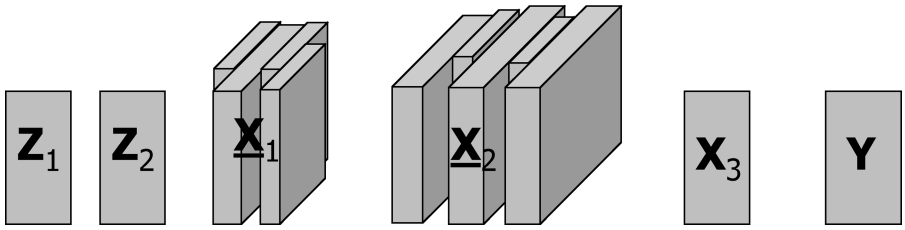

FIG. 7. Processes data may be collected from several unit operations. In this example, two batch processes in sequence ($\underline{\mathbf{X}}_1$, $\underline{\mathbf{X}}_2$) are followed by a continuous process $\mathbf{X}_3$; information is available on raw materials $\mathbf{Z}_1$, and other preprocessing data $\mathbf{Z}_2$; all of these have an effect on the final quality $\mathbf{Y}$.

overall process, as well as scores per block. Hotelling's T^2 and SPE charts can be built for the entire process, and contribution plots are used for fault diagnosis as before. Furthermore, a process design space can be built for the entire process, taking into account all unit interactions.

In a multiblock analysis of a single batch process for example, one could have the combination ($\mathbf{Z}_1$, $\underline{\mathbf{X}}_1$, and $\mathbf{Y}$); $\mathbf{Z}_1$ could include information available on recipes, preprocessing times, hold times, as well as information of the shifts (which operator was in charge). $\underline{\mathbf{X}}_1$ would include process variable trajectories, and $\mathbf{Y}$ 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 (31). This set up would also be used if we wanted to forecast the necessary operating conditions in the batch, so that they correct for raw material variability in a feedforward control mode (see control section discussion).

Several algorithms have been reported for multiblock modeling. The reader can consult (58) where all of the algorithms are presented and compared. As might be expected in multistage processes, there can be significant time delays between the moment an event occurs in one unit (and therefore it 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 traced easily, and eventually one can related a specific lot number to several process stages (via a multiblock PLS). So the process conditions of these units can be used to predict the quality of the product. The reader should be cautious however, because there are situations 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. For example a product may first be produced in 3 parallel reactors (chosen in any combination from a group of four reactors), the out streams mixed and then split into two separators (chosen in any combination from a group of three) for purification; the out streams of the two separators, are given a lot number each. In that case, relating the properties of a product with a specific lot number to the process conditions that the material experienced, is not a straightforward task. A discussion on monitoring difficult multistage operations can be found in (40). In those cases the best alternative in order 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.

One of the early applications of multiblock methods to a pharmaceutical process is that reported by Westerhuis and Coenegracht (60) where they used MB-PLS to model the two-step pharmaceutical process of wet granulation and tableting. Besides the process variables of both steps and the composition variables of the powder mixture, the physical properties of the intermediate granules are also used to model the crushing strength and disintegration time of pharmaceutical tablets. With MBPLS the highly collinear granulate properties can be segregated from the process and composition variables to study separately the influence of both groups of descriptor variables on the tablet properties.

MB-PLS can also be used to break large sets of predictor variables (several types of spectra) into meaningful blocks, and weigh then differently for better prediction. Such an attempt is reported in one of the earliest PAT applications by Frank et al. (61). It is worth noting that the authors emphasize that their objective was to show how to improve the prediction power of the PLS model by various feature selections by incorporating more information (spectral or process variables). Predictor variables from different spectral sources (IR and sonic spectra) are treated as separate blocks, combinations of which, according to a preset pathway, give prediction models for quality-control variables of an industrial product. The interesting part of this work was that the authors used information from the process data to correct the sonic spectra (that is, they performed one form of data laundering).

BATCH TO BATCH MONITORING

The structure of multiblock analysis is also useful for batch to batch monitoring. Data from prior batches may be useful for the analysis and monitoring of future batch processes, if they contain some information on effects that will have an influence on the performance of the future batches. As an example, consider the case where common source of raw materials is being used for successive batches, and the materials from this source have some characteristics (for example, impurity concentrations, surface chemistry properties, and so on) which change slowly with time. Details on this approach can be found in (62).

DESIGN SPACE, PROCESS SIGNATURE

Earlier it was discussed that it is not sufficient to characterize a product with "end point quality measurements." The reason is that the same "measured" quality properties may sometimes be achieved by taking different process paths. Furthermore, for some products we do not measure all the possible quality properties (example, downstream processability). 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 are on target, the properties that determine other characteristics (i.e., the processability of the product) may not be within acceptable limits. Therefore the "process path to the end point" must also be

examined. This "process path to the end point" is also discussed in the European Regulatory Perspective (22) 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).

It should be pointed out here that *the process signature* in the multivariate statistical process control context is nothing else but the two multivariate indices Hotelling's T^2 and SPE. As a matter of fact, these indices take it to account not one feature (e.g., water addition rate or, drying rate) but the combination of all the variables affecting the process and product and their correlations both at each time interval but also their time correlations for the duration of the process (auto and cross correlations for the entire batch). They are therefore a more powerful tool to describe the "*overall process signature*."

Furthermore, these indices can be directly related to the concept of the design space. The "design space" is defined as "the established range of process parameters/formulation attributes that has been demonstrated to provide assurance of quality." Maintaining the T^2 and SPE within their good operation limits (obtained by relating process and raw material variability to acceptable product quality) is nothing more (or less) than ensuring that the operation (raw materials and process) is within the design space.

RAW MATERIAL MULTIVARIATE SPECIFICATIONS

Dushesne and MacGregor (63) presented a methodology for establishing multivariate specification regions on raw/incoming materials or components. 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 from 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 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 **Z**, **X** and **Y**; **Z** contains measurements on *N* lots of raw material data from the past; **X** contains the steady state processing conditions used to process each one of the *N* lots; **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.

E-SENSES: ELECTRONIC NOSE, TONGUE, EARS AND EYES

Apart from the electronic nose technology that has by now become established and that has found its way in pharmaceutical industry (64), other e-sensors are becoming more and more attractive.

The future electronic ear is being developed. Acoustic signals (sound or vibration) from a process may be used for equipment monitoring, as well as process and product monitoring. Acoustic chemometrics, was defined by Esbensen et al. (65): "Acoustic chemometrics is simple: obtaining problem-dependent acoustic signals (by relevant technical means), which—followed by some form of pertinent signal analysis—are subjected to chemometric data analysis. In this context it is often the power of multivariate calibration that comes to the fore."

Examples of the methodology have been reported for fault detection for mechanical equipment (66), for estimating the mass flow by utilizing vibrations from pneumatic conveying of granular materials and for developing a soft sensor to predict the particle-size distribution of a powder (67) from vibrations generated by the powder particles falling onto a surface. In this latter case the ultimate goal was to utilize this in a industrial process where real time vibration measurements will predict product quality.

Bruwer et al. (68) presented an example from the snack food industry. The study involved a series of trials performed on a pilot plant. Several end-use product properties were predicted from the acoustic signals obtained at key locations in the process. The soft sensor enables real-time prediction of product properties that were previously only available via offline laboratory assays.

Electronic eyes are also being developed for detecting defects. Recent work (69–71) on multivariate image analysis (MIA) methods and their extension to online monitoring provides a breakthrough in this area and has led to the use of these imaging sensors for monitoring and controlling industrial processes. Digital imagery makes it possible to monitor solids and other heterogeneous materials (pulp and paper products, polymer films, multiphase streams); it provides informative, inexpensive and robust on-line sensors for the solids industry. Therefore, it opens new ways for the successful monitoring and control of processes, which was traditionally difficult due to lack of sensors. There is potential of utilizing these methods in pharmaceutical industry.

Multivariate image analysis (MIA) describes a set of techniques that employ multivariate statistical methods such as

principal component analysis and partial least squares to analyse images. With MIA most of the analysis is done in the latent variable space rather than the image space. The objective of the approach is to extract subtle information from the image that is related to product quality, and use such information for prediction monitoring and control. This approach is different from the traditional digital image processing, which involves methods for altering the visual image in some way in order to make it more visually appealing or to extract information on the shapes boundaries or location of various observable features. Several applications are listed in (34).

A new machine vision approach for quantitatively estimating and monitoring the appearance and aesthetics of manufactured products is presented by Liu and MacGregor (70). The methodology is specifically designed to treat the stochastic nature of the visual appearance of many manufactured products. This nondeterministic aspect of product appearance has been an obstacle for the success of machine vision in many industries. The emphasis of this approach is on the consistent and quantitative estimation of continuous variations in visual appearance rather than on classification into discrete classes. This allows for the online monitoring and the eventual feedback control of product appearance. This approach is successfully applied to the estimation and monitoring of the aesthetic quality of manufactured stone countertops. This new machine vision approach combines methods such as wavelet texture analysis, latent variables methods for the quantitative estimation of visual quality from textural information, and multivariate statistical process control (MSPC) of visual quality.

The use of chemical multisensor systems or electronic tongues has also been reported for quality monitoring (72, 73). The principle of the electronic tongue is based on utilization of non-specific or low-selective potentiometric chemical sensors with enhanced cross-sensitivity to as many different components in solution as possible. Cross-sensitivity means that the sensor responds not to a single analyte but to several substances simultaneously present in the analysed media. The electronic tongue based on an array of 30 non-specific potentiometric chemical sensors has been applied to qualitative and quantitative monitoring of a batch fermentation process of starting culture for light cheese production. Process control charts were built by using PLS regression and data from fermentations run under "normal" operating conditions.

An alternative form of electronic tasting (sensing how spicy snack food is) is based on MIA and has been used for real-time feedback control; this application is operating with excellent results for several years (74, 75). Inferential sensors based on a digital imaging system have 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 in snack food industry (74, 75). The imaging system is used to monitor these product quality variables and to detect and diagnose operational problems in the plants. It is also used to implement closed-loop feedback control over coating concentration.

PROCESS CONTROL TO ACHIEVE A DESIRED PRODUCT QUALITY

The term "control" currently appears in the PAT 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 insure the output of the process conforms with pertinent specifications.

In this work, terms related to process control, are uses 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 (as for example measurements on product quality).
- 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 (as for example information on raw material).

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 (multi-unit) 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 PAT context in Pharmaceutical industry, where chemometrics was involved, is described by Westerhuis et al. (76). 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, etc. 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 (moisture in the granulation and compression force) based on the intermediate properties, to achieve desirable final properties of the tablets.

To the author's knowledge there are several 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 chemometrics 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 **Z**, **X** (or $\underline{\mathbf{X}}$) and **Y**. A detailed example application can be found in a recent work (Master Thesis, McMaster University, Hamilton, |Ont., Canada, 2006), entitled "Industrial batch data analysis using latent variable methods", by C.P. Rodrigues.

End Point Determination

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

The use of on-line NIR for feedback control in the food industry is described by Osborne (77). NIR on-line samplers were first developed to measure the protein content of flour. This is still the most popular application and provides an excellent example of an NIR feedback control system. Dried gluten is commonly used, particularly in Europe, to replace wholly or partly the protein in flour which would otherwise be derived from high-protein wheat in the grist. The success of on-line NIR for monitoring flour protein content therefore led to its incorporation into a closed-loop control system for gluten addition to flour. A mixing screw is installed between the gluten feed and the NIR sampler station from which a feedback signal controls the gluten feeder. The system has proved to be an efficient and accurate method of control of gluten addition to achieve a target protein content in the flour.

Another example is described by P. Findlay et al. (78), where NIR spectroscopy is used to determine granulation endpoint. The moisture content and particle size determined by the near-infrared monitor correlates 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. There has not been a report of the actual control scheme yet.

Multivariate Manipulation of Process Variables

The question is: Is end point control enough for control? As discussed earlier regulating only the final value of a property (or even several properties) is not 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 "overall" quality, even if the trajectory (path to end point) of one quality variable follows the same desired path in both of 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 but also in scale-up. Latent variable methodology allows for taking into consideration the process variable trajectories in a multivariate way.

Control of batch product quality requires the on-line adjustment of *several* manipulated variable trajectories such as temperature, monomer, initiator or other reactant feed rate trajectories. Traditional approaches based on detailed theoretical models are either based on non-linear differential geometric control or on-line optimization. Many of the schemes suggested in the literature require substantial model knowledge (79) or are computationally intensive and therefore difficult to implement in practice. Empirical modeling offers the advantage of easy model building.

Empirical models for the control of batch product quality where the control action was restricted to only a few movements in the manipulated variables (additional reactant injections) have appeared in the literature and applied in industrial problems. Obviously these approaches can be used in situations where few adjustments are enough to compensate for disturbances and bring the desired end product quality within specifications.

Lately, latent variable methods have found their way to control of batch product quality and have been applied in industrial problems. Latent variable methodology allows for taking into consideration the process signatures in a multivariate way for end point detection problems. Marjanovic et al. (80) describe a preliminary investigation in to the development of a real-time monitoring system for a batch process operated by Aroma and Fine Chemicals Limited. 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 on-line, and initial results indicate that it offers the potential to reduce operating costs.

In another application (81) latent variable methodology was used for soft sensor development that could be used to provide fault detection and isolation capabilities and that it 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 (82). 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 is 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 principal component analysis (PCA) models of the batch processes has been presented. The method presented by Nomikos and MacGregor (36–38) is capable of modeling three-way structures generated when formulating the control problem of batch processes using latent variables.

ESTIMATING BATCH TRAJECTORIES FOR NEW PRODUCT FROM HISTORICAL DATA

The problem that is addressed here is to determine, by utilizing historical data bases, the process operating conditions that will yield a product with a desired set of quality properties. The historical data basis contain information on previous grades and their corresponding process conditions.

Data analysis based formulation of new products was first reported by Moteki and Arai (83), who used principal components analysis (PCA) to analyze data from a polymer production facility. This analysis led them to infer conditions that would result in the production of new polymer grades. Jaeckle and MacGregor (84) used empirical latent variable models which are fitted with historical data from the process. Another methodology was presented (85) where Genetic programming (GP) is used to automatically generate accurate nonlinear models relating latent vectors for the **X** and **Y** blocks.

An extension to the Jaeckle and MacGregor technique, when the operating conditions include a set of time varying profiles for the manipulated variables as in the case of a batch is addressed in (86). The effect of including constraints in the objective function to minimize in order to estimate the desired score vector is analyzed. The multiple solutions contained in the null space are studied to select the final variable trajectories. An industrial case from the pulp and paper industry is used to explain and illustrate the key concepts.

Estimating the Process Trajectories

The methodology in (86) is further extended to include constraints in the process operation itself and additional optimal criteria in the design. The methodology is illustrated 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 (47–49). The cumulative time or, used time, is added as an extra variable trajectory after the alignment of the batches.

RAPID SELECTION OF REPLACEMENT RAW MATERIALS AND MANUFACTURING CONDITIONS

A data-based approach to the development of industrial polymer blends with specified final properties (87, 88) has been presented. The authors discuss that are basically three major degrees of freedom to control the final product properties: the selection of raw materials, the selection of the ratios in which to blend them, and the selection of process conditions used to manufacture them. They present a new optimization approach that simultaneously addresses all of these degrees of freedom, but the primary focus is on the selection of the materials and their ratios. The approach involves building partial least-squares (PLS) models that combine databases on previously made blends and databases on the properties of the component materials used in these blends. The resulting models are then used in an optimization framework to select raw materials from much larger databases (including materials never previously used) and to select the ratios in which to blend them in order to yield a blend product with specified end properties at a minimum cost. The methodology was applied very successfully to two industrial polymer blending problems involving the replacement of materials. The methodology is general and should be applicable to a wide variety of product development problems, such as the development of new catalysts, food products, pharmaceuticals, etc. The methodology is also capable of optimizing over the process conditions to be used for processing the new materials.

PRODUCT TRANSFER AND SCALE-UP

These methods address the following problem: Estimate the operating conditions of plant B in order to produce the same product that is currently produced in plant A, utilizing historical data from both locations from other products. This problem applies to product transfer to different sites and to scale-up.

From the discussions, so far it can be concluded that:

(i) The quality properties of a product should always be checked within a multivariate context, because univariate charts may be deceiving.
(ii) The end-point quality is not sufficient to characterize a product. To achieve consistency in all the product properties (measured quality and ability to process downstream) the process conditions (path to end point) must also be kept in statistical control.
(iii) Process Signatures should therefore be considered; an overall process signature as given by T^2 while SPE and their confidence limits (defined appropriately) give the limits of the design space.

If these points are important for the quality of a product, they should be important for product transfer from site-to-site and for scale-up. 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) while at the same time the overall process signatures have to be mapped.

A methodology for product transfer and scale-up, based on latent variables, is discussed in (89). The methodology utilizes data bases 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.

HOW MANY REAL TIME ANALYZERS DO WE NEED?

There may be a misconception that multivariate analysis requires the installation of many analyzers to generate data in order to uncover correlations in the variables and process signatures. Lange (90) states that: "Another example of PAT Hype is the push by some to install a flurry of instruments in all parts of the manufacturing process. The assumption is that the more data, the better. While it has been correctly stated that measuring end point alone is insufficient to ensure that the process is under control (Lange here references article 91) the alternate extreme is not true; that every process vessel should be equipped with every imaginable sensor."

While multivariate analysis can handle very large data sets, if they are already available, it does not require extra installations. As a matter of fact by utilizing the structure of the existing typical measurements (temperatures, pressures, etc), MSPC can be achieved with *less* real time analyzers. In article (91) it is not suggested adding more instrumentation for more quality measurements. On the contrary, that article suggests to utilize the existing *process measurements*; it cautions the reader that it is not sufficient to rely on the existing on- line quality analyzers, only, but that it should be checked that the process data (i.e., temperatures, pressure, etc.) follow a desired path, utilizing MSPC.

Therefore it is not clear why the author in (90) misinterprets that MSPC requires "every imaginable sensor." As a matter of fact, the experience from other industries in the process systems community is that too many analyzers are not desirable because of maintenance issues. This is the reason that on-line soft sensors are gaining more and more popularity. MSPC acts like an overall soft sensor (5).

Inferential Modeling/Soft Sensors

Accurate on-line measurements of quality variables are essential for the successful monitoring and process control. However, due to measurement difficulties, sometimes process variables may be used to "infer" product quality in real time and therefore replace an analyzer. This is the idea of soft sensors. In many monitoring and control situations we are often lacking real time sensors capable of measuring many of the responses of interest, because the measurement equipment for such quality variables may be very expensive, or difficult to put on-line, or costly to maintain. As a result we often try to develop soft sensors or inferential models which use other readily available on-line measurements such as temperatures, and can be used to infer the properties of interest in a real time manner (92). In a recent paper (81) it was demonstrated through application to a benchmark simulation of a fed-batch fermentation process that mutli-way PLS can provide accurate inference of quality variables, such as biomass concentration, that are often difficult to measure using on-line sensors. It was also demonstrated that the same PLS model can be used to provide early detection and isolation of fault conditions within a fermenter.

The soft sensors can either replace the hardware sensor (analyzer) or be used in parallel with it to provide redundancy and verify whether the hardware sensor is drifting or has failed; when used in parallel the soft sensor will either estimate the property and compare its value with that of the analyser, or it will keep track of the correlation between the analyser reading and the process measurements. An example where a soft sensor is used to assess the reliability of an analyser was presented in (32). Latent variable modeling was used for this purpose.

DATA ACQUISITION AND ARCHIVING

Access to Raw Data from Analyzers

Some analyzer manufacturers do not supply the raw spectral data so that they can be archived together with other data of the process. Rather they supply the calculated quality value that the analyzer is used for. It is important that the plant has access to the raw spectral data, so that can they be used in the future for multivariate analysis to investigate production issues.

Compression

The quality of the data that are archived by the historians is extremely important for reliable results in multivariate analysis. Data acquisition, compression and reconstruction, if not done properly, may distort the quality of the historical data and make them useless for multivariate analysis (and in some cases any other analysis). In order to save storage space, data are compressed; not all data acquired at the designated time intervals are stored. Then the data are reconstructed at the user's request in minute averages, hourly averages, etc. The problem is that *univariate* data compression methods are used to minimize the amount of data that needs to be stored, and they corrupt the multivariate nature of the data. The type and degree of compression should be correctly decided with the vendors. These issues are discussed in detail in (40).

OTHER PRACTICAL CONSIDERATIONS

A detail account of implementing MSPC on line is given in (32), both for continuous and batch processes. There are several issues to consider when implementing MSPC in real time industrial applications. They include the choice of the reference set that will be used for modeling, the choice of variables to include in the monitoring scheme, the choice of weights for the variables, frequency of sampling, to name a few. These issues are discussed in detail in (31, 32) where there is also a discussion of how process related knowledge and other information can

be used in the model and can also help the choice of variable transformations and variable grouping with different weights. An additional issue for batch processes are incomplete data trajectories; the problem has been addressed successfully in (93).

Before any on-line implementation, the model should be tested off line, utilizing past data with faults, to make sure that these faults can be detected. The new monitoring scheme (based on the latent variables) runs for some time period in parallel with the scheme/tool that the operator has been using for the past years. The operators switch to the new scheme after they feel comfortable that they can rely on it. How should the PAT idea be applied in the plant? In small units first? Plant wide? What is the experience with other industries?

Most of the industrial MSPC applications reported in the literature and others known to the author from personal experience, use the methodology to monitor the process performance of specific units, rather than that of the entire plant. The reader is referred to a recent paper written by industrial practitioners (33), where several issues, related to implementing MSPC in industry are discussed in detail based on their experience. They include discussions on data selection and preparation, model development, data preprocessing, filtering, model evaluation. They also discuss how systems are evaluated as candidates for on-line applications and they describe the four stages of on-line implementation (system design, integration, evaluation with on-line systems and maintenance).

CONCLUDING REMARKS

Chemometrics has played an integral part for the real time development of process analytical measurements (multivariate calibration) and it is ready to face the demands of PAT in a wider definition that links PAT to process control, real time process signature monitoring, process understanding and correct technology transfer. In this paper it was demonstrated that multivariate, data-based statistical methods, play a very critical role in providing solutions to these issues. From determining the acceptability of raw material entering the plant to ensuring quality of the product that leaves the plant, the multivariate analysis philosophy should govern all the operations that take that raw material and convert it to a final product in a cost-efficient way, while meeting safety and environmental constraints.

REFERENCES

1. J. B. Callis, D. L. Illman, and B. R. Kowalski, Process analytical chemistry. *Analytical Chemistry* 59 (1987):624A–635A.
2. J. Workman, Jr., M. Koch, and D. Veltkamp, Process analytical chemistry, *Analytical Chemistry* 77 (2005):3789–3806.
3. Guidance for Industry, PAT A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, U. S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Veterinary, Medicine (CVM), Office of Regulatory Affairs (ORA), Pharmaceutical CGMPs, September 2004. ⟨http://www.fda.gov/cder/guidance/ 6419fnl.htm⟩.
4. Pharmaceutical CGMPs for the 21st Century-A Risk-Based Approach, Final Report, September 29, 2004. ⟨http://www.fda.gov/cder⟩.
5. T. Kourti, Process Analytical technology and multivariate statistical process control. Wellness index of process and product—Part 1. *The Journal of Process Analytical Technology* 1, 1 (2004):13–19.
6. J. Workman, Jr., M. Koch, and D. J. Veltkamp, Process analytical chemistry. *Analytical Chemistry* 75, 12 (2003):2859.
7. J. Workman, Jr., K. E. Creasy, S. Doherty, B. Leonard, M. Koch, A. H. Ullman, and D. J. Veltkamp, Process analytical chemistry. *Analytical Chemistry* 73, 12 (2001):2705.
8. J. Workman, Jr., D. J. Veltkamp, S. Dohe, B. B. Anderson, K E. Creasy, M. Koch, J. F. Tatera, A L. Robinson, L. Bond, L. W. Burgess, G. N. Bokerman, A H. Ullman, G. P. Darsey, F. Mozayeni, J. A. Bamberger, and M. S. Greenwood, Process analytical chemistry. *Analytial Chemistry* 71, (12) (1999):121R.
9. W. W. Blaser, R A. Bedeweg, R. S. Harner, M. A. LaPack, A. Leugers, D. P. Martin, R. J. Pell, J. Workman, Jr., and L. G. Wright, *Analytical Chemistry* 67 (1995):47R.
10. K. R. Beebe, W. W. Blaser, R. A. Bredeweg, J. P. Chauvel, Jr., R. S. Harner, M. LaPack, A. Leugers, D. P. Martin, L. G. Wright, and E. D. Yalvac, Process analytical chemistry. *Analytical Chemistry* 65 (1993):199R.
11. B. Lavine and J. J. Workman, Jr., Chemometrics. *Analytical Chemistry* 76 (2004):3365–3372.
12. B. K. Lavine and J. Workman, Jr., Chemometrics. *Analytical Chemistry* 74 (2002):2763–2770.
13. B. K. Lavine, Chemometrics. *Analytical Chemistry* 72 (2000):91R–97R.
14. C. E. Miller, Chemometrics for on-line spectroscopy applications—Theory and practice. *J. Chemometrics* 14 (2000):513–528.
15. R. G. Brereton, Introduction to multivariate calibration in analytical chemistry. *Analyst* 125 (2000):2125–2154.
16. K. Bakeev, *Process Analytical Technology Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries* (Blackwell Pub Professional, 2005).
17. H. W. Siesler,Y. Ozaki, S. Kawata, and H. M. Heise, *Near Infared Spectroscopy Principles, Instruments, Applications*. (Wiley VCH Verlag, Gmbh, 2002).
18. R. Feudale, N. Woody, H. Tan, A. Myles, S. Brown, and J. Ferre, Transfer of multivariate calibration models; A review. *Chemometrics and Intelligent Lab Systems* 64 (2002):181–192.
19. T. Kourti and J. F. MacGregor, Process analysis, monitoring and diagnosis using multivariate projection methods—A tutorial. *Chemometrics and Intelligent Laboratory Systems* 28 (1995):3–21.
20. T. Kourti, Optimization and control. In *Design and Analysis in Chemical Research.* Roy Tranter, ed (Sheffield Analytical Chemistry, CRC Press, Boca Raton, FL, 1999). Ch. 9: 314–364.
21. G. T. Muirhead, Process Analytical Technology at GlaxoSmithKline., Presented at the 19th IFPAC , International Forum Process Analytical Technology, Arlington, VA, Jan 10–13, 2005. (© Gordon T. Muirhead, GlaxoSmithKline, Reproduced with permission).
22. C. Graffner, PAT—European Regulatory Perspective. *The Journal of Process Analytical Technology* 2, 6 (2005):8–11.

23. C. Dushesne, T. Kourti, and F. MacGregor, Application of multivariate methods to start-Up, re-start and grade transition problems. *AICHE J.* 48 (2002):2890–2901.
24. J. V. Kresta, T. E. Marlin, and J. F. MacGregor, Development of inferential process models using PLS. *Computers Chem. Engng*. 18 (1994):597–611.
25. A. J. Burnham, R. Viveros, and J. F. MacGregor, Frameworks for latent variable multivariate regression. *J. Chemometrics* 10 (1996):31–45.
26. I. T. Jolliffe, A Note on the Use of Principal Components in Regression. *Applied Statistics* 31, 3 (1982):300–303.
27. J. M. Sutter, J. H. Kalivas, and P. M. Lang, Which principal components to utilize for principal component regression. *Journal of Chemometrics* 6 (1992):217–225.
28. P. R. C. Nelson, P. A. Taylor, and J. F. MacGregor, Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems* 35 (1996):45–65.
29. F. Arteaga and A. Ferrer, Dealing with missing data in MSPC: Several methods, different interpretations, some examples. *Journal of Chemometrics* 16 (2002):408–418.
30. P. R. C. Nelson, J. F. MacGregor, and P. A. Taylor, The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemometrics and Intelligent Laboratory Systems* 80 (2006):1–12.
31. T. Kourti, Process analysis and abnormal situation detection: From theory to practice. *IEEE control systems* 22, 5 (2002):10–25.
32. T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing* 19 (2005):213–246.
33. I. Miletic, S. Quinn, M. Dudzic, V. Vaculik, and M. Champagne, An industrial perspective on implementing on-line applications of multivariate statistics. *Journal Process Control* 14 (2004):821–836.
34. T. Kourti, Symposium report on abnormal situation detection and projection methods. *Chemometrics and Intelligent Laboratory Systems* 76 (2005):215–220.
35. J. F. MacGregor, T. Kourti, and J. V. Kresta, Multivariate identification: A study of several methods. International Symposium on Advanced Control of Chemical Processes ADCHEM '91 (an IFAC Symposium). Toulouse, France, October 14–16. Proceedings (K. Najim and J. P. Babary, eds.) (1991):369–375.
36. P. Nomikos and J. F. MacGregor, Multivariate SPC charts for monitoring batch processes. *Technometrics* 37, 1 (1995):41–59.
37. P. Nomikos and J. F. MacGregor, Monitoring of batch processes using multi-way principal component analysis. *AIChE Journal* 40, 8 (1994):1361–1375.
38. P. Nomikos and J. F. MacGregor, Multiway partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* 30 (1995):97–108.
39. T. Kourti, Multivariate dynamic data modelling for analysis and statistical process control of batch processes, start–ups and grade transitions. *Journal of Chemometrics* 17 (2003):93–109.
40. T. Kourti, Abnormal situation detection, three way data and projection methods—robust data archiving and modeling for industrial applications. *Annual Reviews in Control* 27, 2 (2003):131–138.
41. T. Kourti, J. Lee, and J. F. MacGregor, Experiences with industrial applications of projection methods for multivariate statistical process control. *Computers and Chemical Engineering* 20 (Suppl.) 1996:S745–S750.
42. D. Neogi and C. E. Schlags, Multivariate statistical analysis of an emulsion batch process. *Industrial Engineering Chemistry Research* 37 (1998):3971–3979.
43. A. Kassidas, J. F. MacGregor, and P. A. Taylor, Synchronization of batch trajectories using dynamic time warping. *AIChE J.* 44 (1998):864–875.
44. P. A. Taylor, Computing and Software Dept., McMaster University, Hamilton, Ontario, Canada, personal communication (May 1998).
45. J. A. Westerhuis, A. Kassidas, T. Kourti, P. A. Taylor, and J. F. MacGregor, On-line synchronization of the trajectories of process variables for monitoring batch processes with varying duration. Presented at the 6th Scandinavian Symposium on Chemometrics, Porsgrunn, Norway, Aug. 15–19, 1999.
46. S. García-Muñoz, T. Kourti, J. F. MacGregor, A. G. Mateos, and G. Murphy, Troubleshooting of an Industrial Batch Process Using Multivariate Methods. *Industrial Engineering Chemistry Research* 42 (2003):3592–3601.
47. S. García-Muñoz, Batch Process improvement using latent variable methods, Ph.D. Thesis, McMaster university, Hamilton, Ontario, Canada (2004).
48. S. García-Muñoz, T. Kourti, and J. F. MacGregor, Ninth Scandinavian Chemometrics Conference, SSC9, in Reykjavic, Iceland, August 21–25, 2005.
49. Salvador Garcia-Munoz, Debashis Neogi, Sanjay Mehta, John MacGregor, and Theodora Kourti, Optimization of Batch Processes using Data Driven Latent Variable Models. *Annual Meeting of the American Institute of Chemical Engineers* Cincinnati , Ohio, October 30–November 4, 2005.
50. G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* 18 (2004):231–241.
51. T. Kourti, P. Nomikos, and J. F. MacGregor, Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control* 5(4) (1995):277–284.
52. C. Duchesne and J. F. MacGregor, Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemometrics and Intelligent Laboratory Systems* 51 (2000):125–137.
53. J. F. MacGregor, T. Kourti, and P. Nomikos, Analysis, Monitoring and Fault Diagnosis of Industrial Processes using Multivariate Statistical Projection Methods. Preprints, 13 World Congress of International Federation of Automatic Control (San Fransisco-USA June30—July 5). Volume M (eds M. Kummel, T. McAvoy, and A. Niemi), (1996), 145–150.
54. P. Nomikos, Detection and diagnosis of abnormal batch operations based on multiway principal component analysis. *ISA Transactions* 35 (1996):259–267.
55. B. Lennox, G. A. Montague, H. G. Hiden, G. Kornfeld, and P. R. Goulding, Process Monitoring of an Industrial Fed-Batch Fermentation. *Biotechnology and Bioengineering* 74(2) (2001):125–135.
56. S. Albert and R.D. Kinley, Multivariate statistical monitoring of batch processes: An industrial case study of fermentation supervision. *Trends in Biotechnology* 19 (2001):53–62.

57. Y. Zhang, M. Dudzic, and V. Vaculik, Integrated monitoring solution to start-up and run-time operations for continuous casting. *Annual Reviews in Control* 27 (2003):141–149.
58. J. A. Westerhuis, T. Kourti, and J. F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics* 12 (1998):301–321.
59. D. J. Taube, Process analytical technology in API: Issues, opportunities and challenges, IFPAC Conference, Scottsdale Arizona, January, 2003.
60. J. A. Westerhuis and P. M. J. Coenegracht, Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics* 11 (1997):379–392.
61. I. E. Frank, J. Feikema, N. Constantine, and B.R. Kowalski, Prediction of product quality from spectral data using the partial least-squares method. *Journal of Chemical Information and Computing Science* 24 (1984):20–24.
62. J. Flores-Cerrillo and J. F. MacGregor, Multivariate monitoring of batch processes using batch-to-batch information. *AIChE Journal* 50, 6 (2004):1219–1228.
63. C. Dushesne and J. F. MacGregor, Establishing multivariate specification regions for incoming materials. *Journal of Quality Technology* 36 (2004):78–94.
64. S. Ohmori, Y. Ohno, T. Makino, and T. Kashihara, Application of an electronic nose system for evaluation of unpleasant odor in coated tablets. *European Journal of Pharmaceutics and Biopharmaceutics* 59, 2 (2005):289–297.
65. K. H. Esbensen, M. Halstensen, L. T. Tonnesen, A. Saudland, J. Svalestuen, S. de Silva, and B. Hope, Acoustic chemometrics—From noise to information, Chemometrics and Intelligent Laboratory Systems 44(1) (1998):61–76.
66. C. C. Tan, N. F. Thornhill, and R. M. Belchamber, Principal component analysis of spectra with application to acoustic emissions from mechanical equipment. *Transaction of Institute Measurement and Control* 24(4) (2002):333–353.
67. M. Halstensen and K. Esbensen, New developments in acoustic chemometric prediction of particle size distribution—The problem is the solution. *Journal of Chemometrics* 14 (2000):463–481.
68. M.-J. Bruwer, J. F. MacGregor, W. Bourg, and C. S. Malvaiz, Multivariate Acoustic Imaging for Product Property Prediction. Presented at AIChE Annual Meeting, Austin, Texas, November 7–12, 2004.
69. M. Bharati and J. F. MacGregor, Softwood lumber grading through on line multivariate image analysis. *Industrial and Engineering Chemistry Research* 42 (2003):5345–5353.
70. J. J. Liu, M. H. Bharati, K. G. Dunn, and J. F. MacGregor, Automatic masking in multivariate image analysis using support vector machines. *Chemometrics and Intelligent Laboratory Systems* 79, 1–2 (2005):42–54.
71. J. J. Liu and J. F. MacGregor, Estimation and monitoring of product aesthetics: application to manufacturing of "engineered stone" countertops. *Machine Vision and Applications* 16, 6 (2006):374–383.
72. Y. Vlasov, A. Legin, and A. Rudnitskaya, Electronic tongues and their analytical application. *Analytical Bioanalytical Chemistry*. 373(3) (2002):136–46.
73. K. Esbensen, D. Kirsanov, A. Legin, A. Rudnitskaya, J. Mortensen, J. Pedersen, L. Vognsen, S. Makarychev-Mikhailov, and Y. Vlasov, Fermentation monitoring using multisensor systems: feasibility study of the electronic tongue. *Analytical Bioanalytical Chemistry* 378 (2004):391–395.
74. H. Yu, J. F. MacGregor, G. Haarsma, and W. Bourg, Digital Imaging for On-line Monitoring and Control of Industrial Snack Food Processes. *Industrial and Engineering Chemistry Research* 42 (2003):3036–3044.
75. H. Yu and J. F. MacGregor, Multivariate image analysis and regression for prediction of coating content and distribution in the production of snack foods. *Chemometrics and Intelligent Laboratory Systems* 67 (2003):125–144.
76. J. A. Westerhuis, P. M. J. Coenegracht, and F. L. Coenraad, Multivariate modelling of the tablet manufacturing process with wet granulation for tablet optimization and in-process control. *International Journal of Pharmaceutical* 156 (1997):109–117.
77. B. G. Osborne, Near-infrared spectroscopy in food analysis, *in Encyclopedia of Analytical Chemistry* (ed. R. A. Meyers) (John Wiley & Sons Ltd, Chichester, 2000).
78. P. Findlay, K. Morris, and D. Kildsig, PAT in fluid bed granulation. Presented at AICHE, San Francisco, CA. 2003.
79. D. Shi, P. Mhaskar, N. H. El-Farra, and P. D. Christofides, Predictive control of crystal size distribution in protein crystallization. *Nanotechnology* 16 (2005):S562–S574.
80. O. Marjanovic, B. Lennox, D. Sandoz, K. Smith, and M. Crofts, Real-time monitoring of an industrial batch process. Presented at CPC7: Chemical Process Control, Lake Louise Alberta, Canada, Jan 8–13, 2006.
81. H. Zhang and B. Lennox, Integrated condition monitoring and control of fed-batch fermentation processes. *Journal of Process Control* 14 (2004):41–50.
82. J. Flores-Cerrillo and J. F. MacGregor, Control of batch product quality by trajectory manipulation using latent variable models. *Journal of Process Control* 14 (2004):539–553.
83. Y. Moteki and Y. Arai, Operation planning and quality design of a polymer process, in ZFAC DYCORD (Pergamon Press, Bournemouth, UK, 1986), 159–165.
84. C. M. Jaeckle and J. F. MacGregor, Product design through multivariate statistical analysis of process data. *AICHE Journal* 44, 5 (1998):1105–1118.
85. S. Lakshminarayanan, H. Fujii, B. Grosman, E. Dassau, and D. R. Lewin, New product design via analysis of historical databases. *Computers and Chemical Engineering* 24 (2000):671–676.
86. S. García-Muñoz, T. Kourti, J. F. MacGregor, F. Apruzzece, and M. Champagne. Optimization of batch operating policies, Part I—handling multiple solutions. Industrial and Engineering Chemistry Research; 2006; ASAP Web Release Date: 10-Oct-2006; (Article) DOI: 10.1021/ie060314g.
87. K. Muteki, J. F. MacGregor, and T. Ueda, The rapid development of new polymer blends: The optimal selection of materials and blend ratios. *Industrial Engineering Chemical Research* 2006, in press.
88. K. Muteki, J. F. MacGregor, and T. Ueda, Mixture designs and models for the simultaneous selection of ingredients and their ratios. *Chemometrics and Intelligent Laboratory Systems*, submitted, 2006.
89. S. García-Muñoz, J. F. MacGregor, and T. Kourti, Product Transfer between Sites using Joint Y_PLS. Chemometrics and Intelligent Laboratory Systems, 79 (2005):101–114.

90. A. J. Lange, Guarding against PAT hype. *The Journal of Process Analytical Technology* 2, 3 (2005):18–20.
91. T. Kourti, Process analytical technology and multivariate statistical process control. Index of wellness of product and process—Part 2. *The Journal of Process Analytical Technology* 2, 1 (2005):24–28.
92. S. Park and C. Han, A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. *Computers and Chemical Engineering* 24, 2 (2000):871–877.
93. S. García-Muñoz, J. F. MacGregor, and T. Kourti, Model predictive monitoring for batch processes with multivariate methods. *Industrial and Engineering Chemistry Research* 43 (2004):5929–5941.