2.10 Batch Process Modeling and MSPC

S. Wold, Umeå University, Umeå, Sweden

N. Kettaneh-Wold, Umetrics Inc., Kinnelon, NJ, USA

J. F. MacGregor, McMaster University, Hamilton, ON, Canada

K. G. Dunn, ProSensus Inc., Hamilton, ON, Canada

2.10.1 Introduction

Batch processes operate for a finite period with a defined starting point and ending point and a time-varying behavior over the operating period. They are common in many industries. In traditional specialty chemical industries, they are used for carrying out chemical reactions of various compounds to produce specialty polymers, agricultural chemicals, catalysts, and so forth. They are the mainstay of the pharmaceutical industry where they

are used for the batch production of active pharmaceutical ingredient (API) and other excipients and for the processing of materials into tablets using batch processes such as granulation, drying, and pressing. Fermentation and other cell-cultivation processes in biotech production, as well as the subsequent harvesting and purification are all batch processes. The food industry has many batch processes ranging from mixing and blending to the aging of cheese. One of the most intensive users is the microelectronics industry, where essentially all of the sequential steps of lithography, vapor deposition, etching, and so forth, are batch processes. Most parts manufacturing processes, such as stamping, injection molding, and machining, are batch in nature. Overall, there are a large number of nontraditional areas where the data collected are batch in nature. Examples are transient data obtained from studies on pharmacokinetics, metabonomics, and contrast-enhanced medical imaging.

Particular problems discussed are

1. data preprocessing, including alignment, centering, and scaling;
2. choice of model type and modeling approach;
3. classification of a batch as acceptable or bad on the basis of its process data;
4. identifying variables related to acceptable and unacceptable batches;
5. online prediction of process variable trajectories, optimal end point, and final product quality before completion of the batch; and
6. control of final product quality.

2.10.1.1 Nature of Batch Data

The data available on batch processes fall into three general categories:

1. The initial conditions (**Z**) available before the start of the batch, including raw material properties;
2. The time-varying process variables and online analytical sensor responses measured throughout the batch evolution (trajectories – **X**);
3. The quality of final product and productivity variables (**Y**), usually measured after the completion of the batch.

In general, each of these data types are blocked in matrices as illustrated in **Figure 1**. Each matrix/array has I rows (layers in **X**) corresponding to I batches.

Each of these blocks may be further broken up into several blocks on the basis of the nature of the data, or owing to the presence of several sequential batch stages or several distinct batch operations. A unique characteristic of batch data is the three-dimensional (3D) **X** array containing the time-varying trajectory measurements on all the process variables and online sensors from the start to the end of the batch. Given the presence of these time-varying trajectory data, the number of observations available in a typical batch analysis problem can easily exceed half a million. This dimensionality problem arising because of the presence of the very large **X** arrays, and the fact that there is extreme temporal and contemporaneous correlation among all the variables, provides a strong motivation for using multivariate chemometric approaches for analyzing these data. These approaches project the information in the data onto reduced dimensional latent variable spaces.

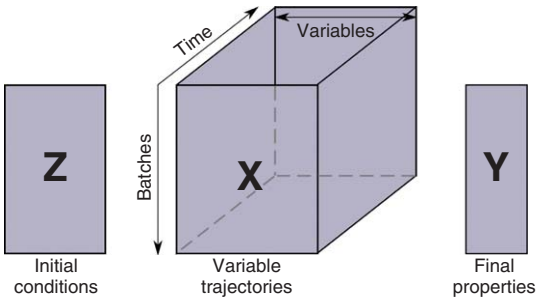

Figure 1 Nature of batch-training set data. Each additional prediction-set batch constitutes one additional row in **Z** and one additional layer in **X**. The latter has to begin with only one vector at time = 1 and fills out with one vector per time point during the evolution of the batch. The final quality (**Y**) is usually based on later laboratory analysis of a sample of the product.

2.10.1.2 Objectives for Analyzing Batch Data

There are four major objectives for analyzing batch data, and this chapter is organized according to these.

1. The analysis of historical batch operations to gain process understanding and to troubleshoot earlier process operating problems. This is usually the first step in any investigation of batch processes and hence precedes subsequent investigations.
2. The monitoring or statistical process control (SPC) of ongoing batch operations.
3. The prediction of final product quality, or end point, on the basis of evolving batch data.
4. The optimization and/or active control of batch operating conditions to achieve a desired final product quality.

2.10.2 Industrial Examples

In this chapter, we briefly describe three industrial batch processes and the type of data available from them. They are used at various points throughout the remainder of this chapter to illustrate different approaches for batch-process modeling, monitoring, prediction, and control.

2.10.2.1 A Batch Drying/Reacting Process in the Manufacture of an Agricultural Chemical

This batch process is a critical step in the manufacture of an agricultural chemical. A mass of wet cake is charged to the batch vessel to evaporate and collect the solvent contained in the initial charge. Some important chemical/structural modifications also occur during this drying stage. The mass of the cake and several other chemistry measurements from the cake are available before the drying process starts (**Z**). The drier is started and the solvent is collected in an external tank, while temperature and agitator profiles are measured and adjusted to obtain the desired final product properties (**Figure 2**). The final properties (**Y**) are assessed through eight measurements.

The process can be thought of as being divided into three stages (phases): stage 1 runs from the beginning to the point when the agitator is turned up to high speed, stage 2 is from this point until the time when the temperature reaches its maximum, and stage 3 is from this point until the end of the batch (**Figure 2**). Further details are in the paper by Garcia-Muñoz *et al.*[1]

A principal components analysis (PCA) on the eight critical **Y** measurements from 59 batches showed clear separation between the on-specification and off-specification batches (**Figure 3**). The subsequent sections will analyze how the **Z** (initial conditions) and **X** (trajectory) data are related to poor quality

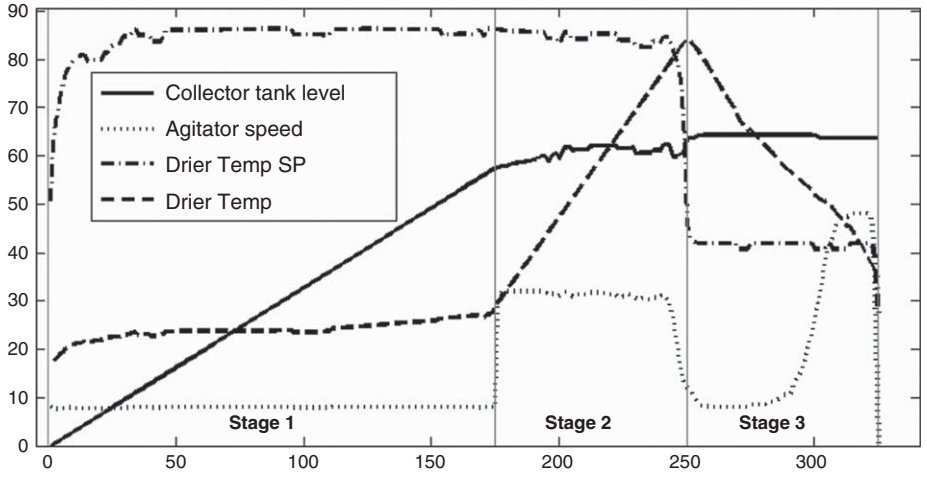

Figure 2 Selected trajectories for a batch from the drying data set.

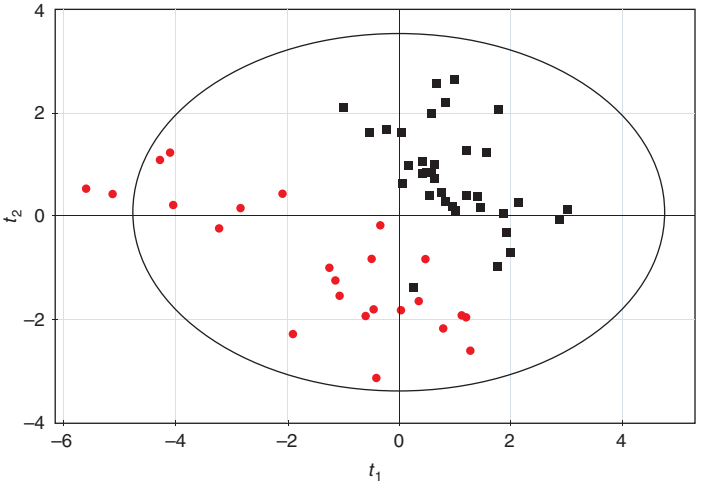

Figure 3 The PCA scores from the eight **Y**-space measurements show a clear separation in the on-specification (squares) and off-specification batches (circles).

product and how problems in the initial conditions and trajectories can be easily detected and fixed, thereby leading to improved batch operations.

2.10.2.2 Nylon Autoclave Reactor

These data on an industrial nylon autoclave reactor have been used in the batch analysis and monitoring literature[2] and are readily available (http://www.prosensus.ca/white-papers/). They consist of trajectories (249 time points) on 10 process variables for 61 batches. See **Figure 4** for the trajectories from one batch. Further description of the data and some results of their analysis are given by Kourti *et al.*[2] and so are not repeated here. In this chapter, we use these data to illustrate batch monitoring and the prediction of future trajectories at any period during the batch.

2.10.2.3 Fermentation Process for Yeast Production

Data from 33 batches of the last (fifth) step of baker's yeast production in the year 1999 at Jästbolaget AB, Solna, Sweden, were obtained in a feasibility study of multivariate batch process SPC (MSPC). $K = 7$ process variables

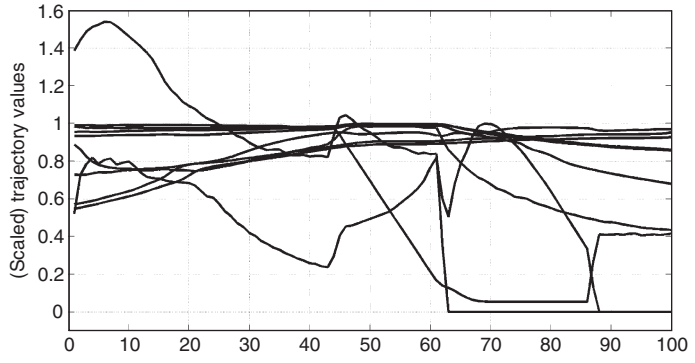

Figure 4 Nylon batch autoclave data: Trajectories for one batch showing the $K = 10$ process variables. Variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_6$, and $\mathbf{x}_7$ are temperatures; variables $\mathbf{x}_4$, $\mathbf{x}_8$, and $\mathbf{x}_9$ are pressures; and variables $\mathbf{x}_5$ and $\mathbf{x}_{10}$ are flows.

Table 1 Batches of the yeast training and prediction sets

Training set, PCA	B, C, I, M, N, Q, R, T, V, X, Z, a, b_1, c_1, d, e, f, g, h, i_1
Prediction set, PCA	A, G, H, J, L, O, P, U, Y, D, j, k, l

Three training set batches (B, C, V) have no yield values and are hence not part of the partial least squares (PLS) training set for the batch model. The four good batches in the prediction set are D, j, k, and l.

were measured every 10 min over 14 h to give $\mathcal{J} = 83$ time points for each batch. The variables were ethanol content, temperature, molasses feed, NH_3 feed, air flow, tank level, and pH. In addition, for each batch, one single value of the inoculate (initial condition, **z**) and one result variable (yield, **y**) were given.

A training set of 20 acceptable batches and a prediction set of 13 batches (4 acceptable and 9 nonacceptable) were formed from the data (**Table 1**). Further details are given by Eriksson *et al.*[3] (Chapter 13).

These data will be used as illustrations of batch process monitoring, and the classification of the final batch as acceptable or not, together with the quantitative prediction of the final yield from the process data plus the inoculate value.

2.10.3 Analyzing Historical Batch Process Data

Analyzing historical batch data is an important step in gaining an understanding about the nature of the effects of varying initial conditions and process operating trajectories on the performance of the batches and on the final product quality. The understanding gained and the models developed at this stage can be used to isolate and diagnose poor operation periods in the past, to define an SPC space for the subsequent monitoring phase, and to obtain prediction and control models.

There are several different approaches for analyzing batch data. In most cases, the differences revolve not only around the objectives of the analysis but also around how to treat the 3D array (**X**) of process operating trajectories. The important information contained in these time-varying trajectories must be extracted through some form of data analysis. The simplest approach is to extract various meaningful landmark features from the trajectories based on our best current understanding of what features might be most influential on the process performance. More sophisticated multivariate chemometric methods (based on PCA and partial least squares (PLS)) provide potentially more powerful approaches to summarizing these process data in terms of a small number of latent variables.

2.10.3.1 Landmark Feature Approach

This approach attempts to capture the essential information in the **X** array about the evolution of the batches by defining landmark features of the process variable trajectories and recording the values of these features for each batch run.

2.10.3.1.1 Landmark approach for example 1, drying data

Referring to **Figure 2** for the agricultural chemical batch drying process, $\mathcal{J}$ landmark features were selected and calculated from the trajectories for each batch. **Table 2** shows the $\mathcal{J} = 13$ landmark features used.

The observed values of these landmark features for each individual batch are then entered as a row in a new **X** matrix, along with the corresponding initial conditions (row of **Z**) and final quality (row of **Y**) for each batch. The resulting data for I batches consist of the $I \times N$ initial condition matrix **Z**, the $I \times \mathcal{J}$ landmark feature matrix **X**, and the $I \times M$ quality matrix **Y** as shown in **Figure 5**.

These data are then analyzed either as a regular PLS (combining **Z** and **X** into a single matrix with proper scaling) or as a multiblock PLS.[2,4] For the agricultural chemical process, a standard PLS analysis gave two components, significant by cross-validation (CV),[5] with the resulting $R^2Y = 50.1\%$ and $Q^2Y = 40.0\%$.

The VIP plot (**Figure 6**) highlights those raw material property (**Z**) variables and landmark (**X**) variables important to the PLS model. This plot shows that the most important variables are from the process operation

Table 2 The landmark features calculated from each trajectory during a specific phase

Name-phase	*Calculated feature*	*Phase*
Level-Mean-3	Mean collector tank level	3
DiffPress-Mean-1	Average differential pressure	1
DryPress-Mean-1	Average drier pressure	1
Power-Slope-1	Slope of the power trajectory	1
Torque-Slope-1	Slope of the torque trajectory	1
DTemp-SP-Mean-3	Average drier temperature set point	3
DTemp-Mean-1	Average drier temperature	1
DTemp-Slope-2	Drier temperature slope	2
Temp1	Drier temperature at the end of phase 1	1
Temp2	Drier temperature at the end of phase 2	2
Time1	Time duration of phase 1	1
Time2	Time duration of phase 2	2
Time3	Time duration of phase 3	3

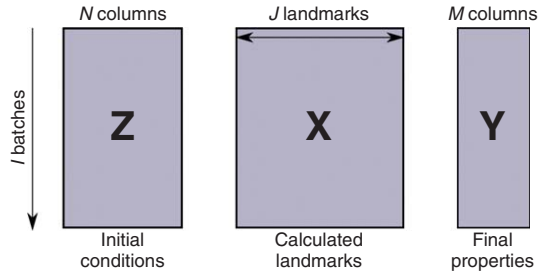

Figure 5 The data structure (of the training set) for the landmark approach: The initial conditions in **Z**, a total of *J* landmark features in **X**, and the *M* final product quality measurements. There are *I* rows, one row per batch. Prediction-set batches have one complete row of **Z** and one row of evolving landmark features in **X**, but no **Y**-vector.

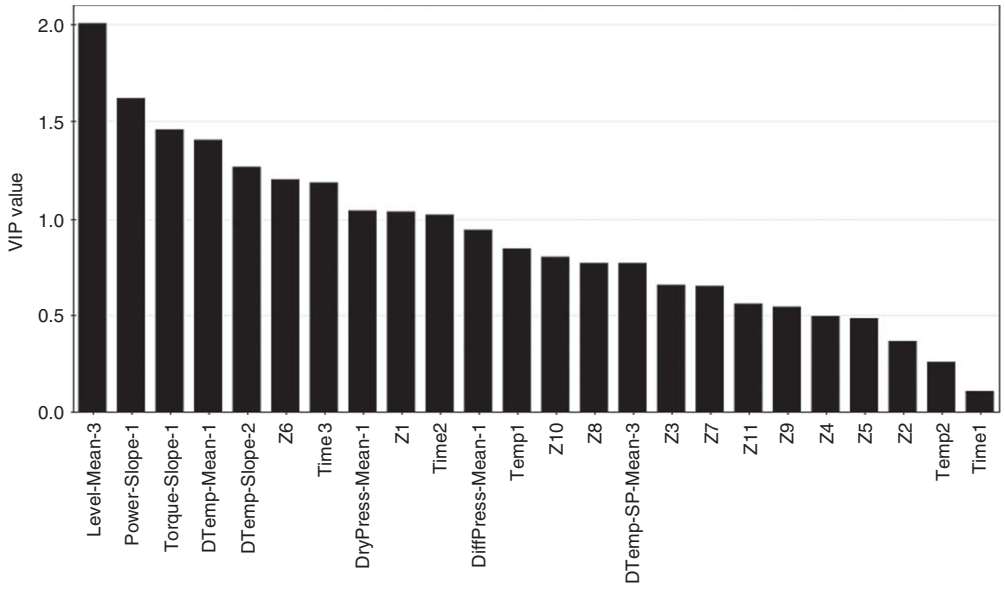

Figure 6 The VIP plot for the PLS model relating the initial chemistry and landmark feature information to the final product quality.

and not from the initial chemistry (**z**'s). This was an important revelation because the company had been focusing on the chemistry variables as the expected cause of poor product.

The score plot in **Figure 7** shows a clear separation between the good and the bad batches. Good batches generally have simultaneously high t_1 and t_2 values. The loading plot (w_1^* versus w_2^*) in **Figure 8** indicates process variables that are related to a large positive t_1 and t_2 values (good quality batches). Notice again that the initial chemistry variables (open circles) have much lower weights than the process variables (especially in the direction of improving quality), indicating that process conditions play a much greater role than the chemistry conditions in determining the process outcome. The plot shows that high rates of increase (slopes) of power, torque, and temperature in phase 1, as well as low solvent level in the collector tank in phase 3 (equivalent to low initial solvent in the wet cake) all relate to a good batch outcomes. All of these conditions could easily be enforced in new batches, thereby ensuring a high percentage of good batches in the future. The only reason this had not been done earlier was that an efficient analysis of the historical batch data had not been carried out to reveal these issues.

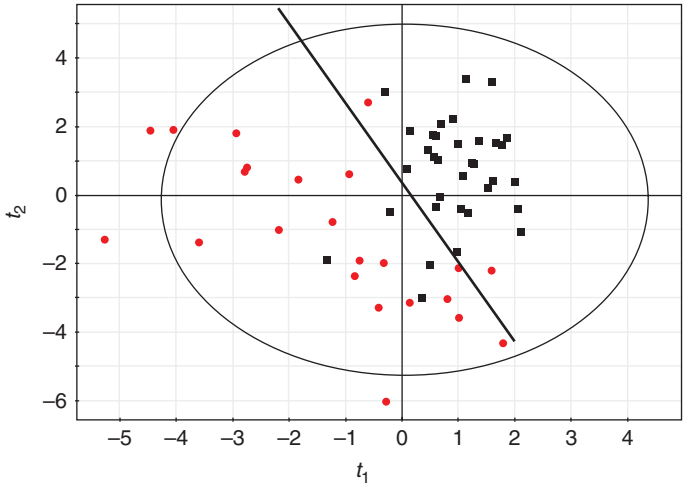

Figure 7 The PLS score plot for on-specification (squares) and off-specification (circles) batches. A fairly clear separation is noticeable, with on-specification batches showing positive values of t_1.

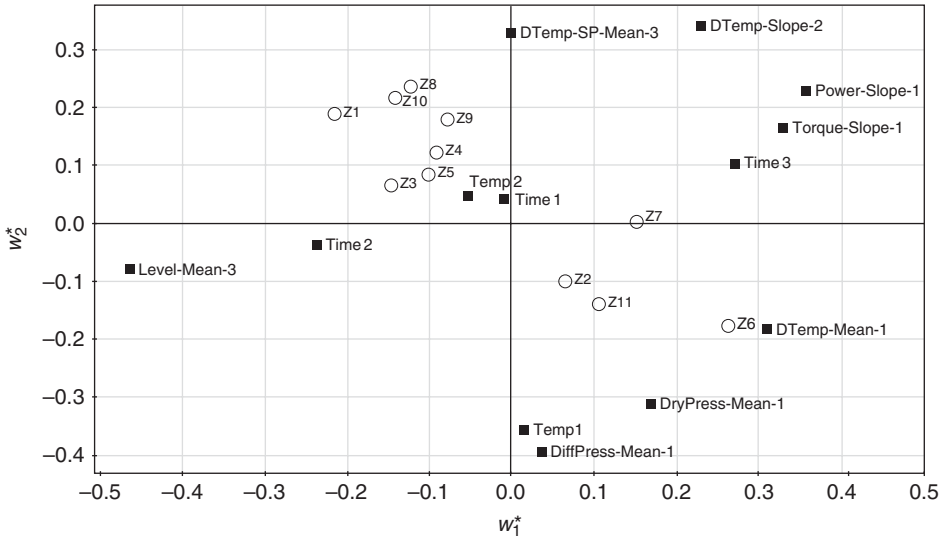

Figure 8 The w^* loadings plot for the chemistry initial condition variables (open circles) and process operating conditions (landmarks) calculated from the trajectories (squares).

2.10.3.1.2 Discussion of landmark feature approach

The landmark feature approach is the simplest approach, both conceptually and technically, and hence should be considered at least in the initial stages of any investigation. However, there can be some difficulties. One is that a sufficient set of landmark features may not be easily extractable from the trajectories. For example, if the trajectories are mostly varying smoothly over time, distinct features may not be obvious. The drier data discussed above were ideal for a landmark feature approach because it had many clearly defined phases and sudden feature changes in the trajectories (**Figure 2**). In contrast, the nylon polymerization trajectories are mainly smooth (**Figure 4**), and there are few obvious landmark features. In such cases, whether a sufficiently predictive set of features can be extracted depends very much on the prior knowledge of the engineer. In many cases, important but subtle features may be overlooked. Another potential problem is that some landmarks may disappear in abnormal batches.

In general, the landmark approach is well suited for batch processes with many distinct operational changes that vary from batch to batch. Examples are the batch blending of components into a final product such as in food or cosmetics manufacturing.

2.10.3.2 PCA/PLS Modeling Approaches Based on Unfolding the X Array

In these approaches, the three-way data array **X** is unfolded in one of two ways, the choice of which depends on the focus of the investigation as well as on the nature of the data.

One approach is to unfold the data batchwise (batchwise unfolding, BWU), that is, to unfold the **X** array such that all the information for each batch is contained in one row as illustrated in **Figure 9(a)**.[6–8] The data are then mean centered and scaled for each column. In this way, the mean centering removes the mean trajectories of each variable, and the variation remaining in the mean-centered **X** matrix is the variation of all the variables about their mean trajectories. Performing any subsequent PCA or PLS analysis on this matrix then summarizes the major sources of variation among the different batches and allows efficient batch-to-batch comparison. This approach also allows for incorporating initial conditions (**Z**) and final product quality variables (**Y**) associated with each batch when performing either a PCA or a PLS analysis.

A second approach unfolds the variables observationwise (observationwise unfolding, OWU), that is, with each row corresponding to an observation at some time in each batch and each column corresponding to the variables measured as illustrated in **Figure 9(b)**.[6,9] Mean centering by column (variable) then simply centers the origin of each variable about zero, but does not remove the average trajectories. The variation remaining is, therefore, the total trajectory variation for each variable. Performing PCA or PLS (using local batch time or a maturity variable as **y**) on these OWU data finds a smaller number of components that summarize the major behavior of the complete trajectories of the original variables. It does not initially focus directly on the differences among batches as BWU does, but on summarizing the variable trajectories.

For analyzing differences among batches, the resulting summary, the score matrix **T**, is then used in a second model, which is similar to the single model of the BWU approach, except that it is based on the unfolded score matrix **T** of the OWU data instead of the BWU raw data. We shall henceforth refer to this double analysis as OWU–TBWU when a score matrix, **T**, is used in the second stage.

The nature of the variation left in the unfolded **X** matrix by BWU and by OWU is illustrated with one of the scaled temperature variables from the nylon batch polymerization data in **Figure 10**. **Figure 10(a)** shows the raw trajectory data for the temperature trajectory for all the batches. **Figure 10(b)** shows the same trajectories after mean centering and autoscaling of the OWU data. Note that the trajectories are identical to **Figure 10(a)** but for now being centered about zero. **Figure 10(c)** shows the same trajectories after unfolding batchwise and mean centering and autoscaling. The variation shown is the deviation of the trajectories about the mean trajectory.

This illustrates one of the fundamental differences in the unfolding methods. PCA or PLS (using local batch time or a maturity variable as the **y**) of the OWU data will find components that best approximate the raw trajectories. PCA or PLS (using final quality **Y**) of the BWU data will find components that best explain the deviations among the batches. The information extracted in the two cases is therefore very different. In the BWU and mean centering, as the average trajectories have been removed, they do not

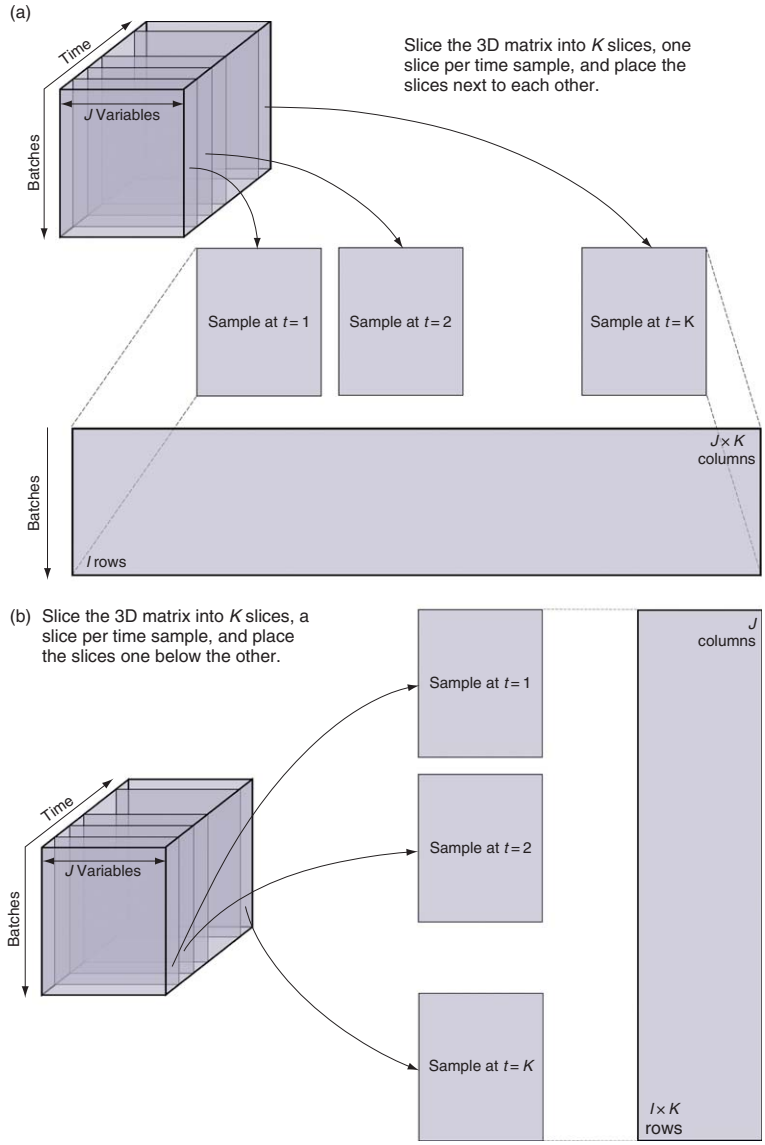

Figure 9 Unfolding the batch data: (a) batchwise unfolding (BWU) and (b) observationwise unfolding (OWU). Note that an equivalent BWU-X matrix in (a) is obtained when the matrix is sliced in the time direction and the $I \times K$ time histories for each of the variables over all batches are spliced into an $I \times J$ K-matrix. An equivalent OWU-X matrix in (b) is also obtained when each $I \times J$ batch matrix is placed under each other. In both cases, the only difference is that the columns or rows are in another order than the ones shown for **X** above in (a) and (b), respectively.

have to be modeled by the PCA or PLS. In **Figure 10**, the variances of the original and the mean-centered OWU data before scaling are the same (715) (**Figures 10(a) and 10(b)**). Upon removing the average trajectory from the data with BWU (**Figure 10(c)**), the remaining variance among batches before scaling is only 9.6.

However, with OWU, after sufficient PCA or PLS components, the OWU residuals are of comparable size to those after zero components of the BWU model (**Figure 10(g)**). This is because the first set of OWU-PLS scores pick up the average trajectories of the highly correlated variables. If there are groups of variables that are correlated in different ways at different time points throughout the batch, several scores are needed to describe the average trajectories of the groups (**Figures 10(d)–10(g)**).

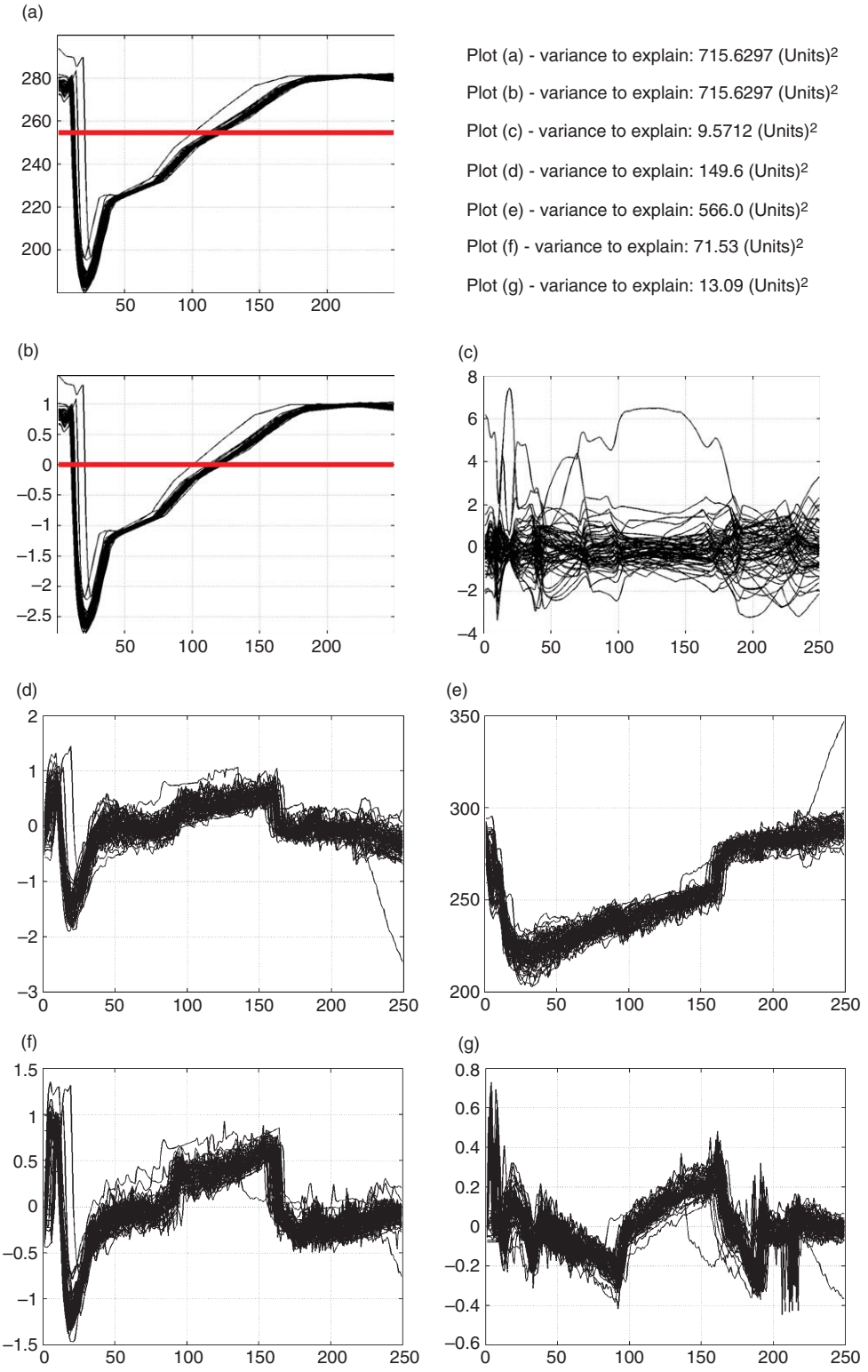

Figure 10 Temperature trajectories from the nylon example: (a) noncentered raw trajectories, (b) centered trajectories, (c) trajectories after BWU and autoscaling, (d) residuals after OWU and one PLS component, (e) predicted values by PLS model of OWU after one PLS component, (f) residuals after OWU with two PLS components, and (g) residuals after OWU with eight PLS components.

2.10.3.2.1 Combining OWU and TBWU (OWU–TBWU)

For analyzing batch data, the main motivation for the OWU–TBWU approach is dimensionality reduction before the BWU stage, which is warranted when the number of variables measured on the process is large. One example is when spectroscopic or chromatographic data are measured on samples taken from the process at regular intervals. The second motivation is to provide a simple basis for online batch process monitoring (B-SPC, see Section 2.10.5.2).

As briefly described above, the batch data array **X** is first unfolded as shown in **Figure 9(b)** (using OWU) to give an $I \times K$ row and $\mathcal{J}$ column **X**-matrix ($\mathbf{X}_{\mathrm{OWU}}$), which is centered and scaled. A guiding **y**-variable (centered and scaled) is added, and a PLS analysis is carried out. The guiding y is preferably a maturity variable (see Section 2.10.3.2.2), or if such is not available, local batch time is used. This is, for each process point, the time from the batch start, and hence runs from 0 to I, and restarts at zero at the beginning of next batch. CV[5] is used to estimate the number of significant PLS components. However, as it is essential that sufficient components be used to well approximate **X**, special rules are used in addition to the standard CV rules to ensure the adequate approximation of $\mathbf{X}_{\mathrm{OWU}}$.

The modeling of the OWU data by a single PLS model works well if the data have the same covariance structure throughout the whole batch evolution. If not, many more components are needed to capture the changing covariance structure, and the models become unwieldy. In the case of changing covariance structure – the typical case – it is practical to divide the batches into phases in such a way that the covariance structure is almost constant, or at least possible to model, within each phase. The ensemble of PLS models, one per phase, will then adequately capture the batch dynamics, and the resulting scores in each phase model, $\mathbf{T}_{\mathrm{g,OWU}}$, will then carry this dynamics to the upper batch level (Section 2.10.3.2.2). This is further discussed in Section 2.10.3.2.5(iv).

In summary, if the process data are divided into phases with fairly homogeneous variance–covariance structures, and these phases are modeled separately with sufficient numbers of components, the batch process dynamics are adequately captured also by the OWU–TBWU approach (see Section 2.10.3.2.6), and the two approaches OWU–TBWU and BWU will be essentially equivalent.

2.10.3.2.2 Alignment of trajectories based on time or a maturity variable

In some batch processes (e.g., the nylon autoclave example), there is very little variation in the total duration of each batch. The batch is programmed to proceed according to a defined schedule (true also for most manufacturing operations of parts).

In other processes, however, the total time to the completion of the batch is different from batch to batch. This is usually due to the batch duration being a variable that is adjusted to achieve an end point based on some measured or inferred process variable. The duration may change because of different sizes of batch charges, different impurity levels that affect the rate of reaction, or cooling rates that vary according to the season, thereby affecting the rate at which a reaction can be carried out. With a process divided into phases, these are rarely prolonged and shortened equally. Hence, phases need to be aligned separately.

The objective of aligning the batch trajectories or the OWU scores is to ensure that, to the extent possible, the variables or scores at any point during one batch correspond to those at the same state in other batches, so that the variables in the columns of the BWU matrix (**Figure 9(a)**) correspond as much as possible. For the OWU–TBWU approach, the data do not need alignment before the OWU modeling, but certainly the resulting OWU scores, **T**, need alignment before the subsequent batch modeling using the unfolded scores. Hence, the discussion below relates to *both* the BWU of raw data and the analysis of the scores of the OWU analysis (batch-level analysis).

There are several possible approaches to the alignment of batch data.[6,10,11] The simplest, if there is very little difference in the batch lengths, is to simply extrapolate or truncate batches to make them the same length. Clearly, this will be unacceptable if there are large differences in duration among batches.

Another is to force each batch to have the same number of points by linearly dividing the batches up into an equal number of intervals and resampling. This assumes that the inequality in duration is linearly distributed throughout every period of the batch. In order not to lose the information on batch duration, the duration of each batch is usually then added as a variable in the **Z** matrix (**Figure 1**).

By far the most powerful approach[6] used widely in industry is to find an indicator or maturity variable, such as the conversion of a key reactant in a reaction. This must monotonically increase (or decrease) from a given

starting point (e.g., 0% conversion) to a final end point (e.g., 100% conversion). This indicator (maturity) variable is then used in place of time, and all variables are sampled in equal intervals of this variable (e.g., at every 1% of conversion). Different maturity variables can be used in different phases. In order that information on the time to reach each value of the indicator variable is not lost, this time usage is also added as a new variable to the **X** array and its trajectory treated in the same way as the other measured trajectory variables. Therefore, the number of variables has not been reduced by use of the indicator or maturity variable; rather the time usage variable has just replaced the maturity variable, and all other variables including the time usage variables are now plotted against the maturity variable.

Some combination of the above approaches usually can handle the alignment of most batch processes. However, if these are not possible, one can always perform dynamic time warping methods[10] that will nonlinearly warp all batch trajectories to match as closely as possible that of a reference batch.

The agricultural batch drying process is an example of a case where the batch durations were widely varying. To align them, a combination of the maturity variable approach (in the first two phases) and linear time stretching (phase 3) were used. **Figure 11** shows some typical unaligned trajectories and their resulting aligned

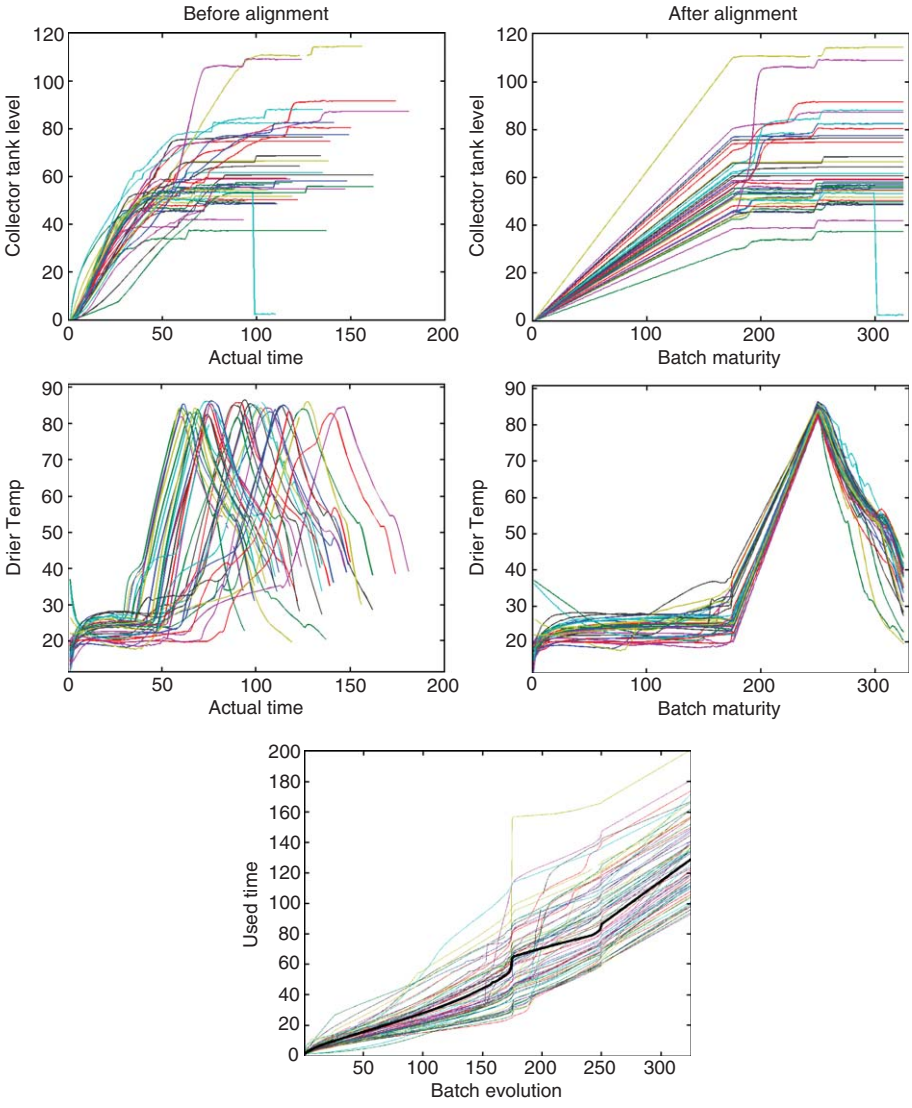

Figure 11 Top – four trajectories: before and after alignment for two of the batch tags. Bottom – the time-usage trajectories.

versions – see Garcia-Muñoz *et al.*[1] for details. Also shown in **Figure 11** is a plot of the time usage trajectories for each batch computed from the alignment. These trajectories show the relative rate at which each batch is progressing relative to all the other batches at every time point throughout the duration of the batches.[1] This variable is very important, as it preserves the timing information from each batch, thereby allowing reconstruction of the original trajectories, and it allows one to relate the rate of progress of any batch during any time interval to the performance and final product quality of the batch (this is illustrated in the analysis of the batch drier example in Section 2.10.3.2.3).

2.10.3.2.3 PCA and PLS analysis of BWU and OWU–TBWU trajectories

With the batch data aligned and unfolded batchwise as shown in **Figure 12**, one can then perform PCA or PLS on any of the blocks, or PLS on the entire multiblock structure. Note that the same analysis is made when the OWU scores are used for the characterization of the batches instead of the raw **X**-data. Therefore, in the remaining sections of this chapter, BWU will be used to describe either BWU with the raw variables or BWU with the summary scores ($\mathbf{T}_{\mathrm{OWU}}$) from OWU. A PCA or PLS on the unfolded **X** or $\mathbf{T}_{\mathrm{OWU}}$ matrix will, therefore, model the covariance structure among the variable (or score) deviations about their average trajectories at each interval over the duration of the batch. This model will capture essentially all of the nonlinear, time-varying behavior in the batch data, first by removing the average trajectories and then by having loadings for each deviation variable at each time interval throughout the batch. Thus, the model provides a locally linearized, time-varying model over the entire batch duration (this is illustrated below on the drier batch data).

A multiblock PLS analysis[2] on the **Z**, **X**, and **Y** matrices of the drier data set showed that three PLS components were sufficient to capture most of the variation as judged by CV ($R^2_{\mathrm{Z}} = 37.8$, $R^2_{\mathrm{X}} = 37.7$, $R^2_{\mathrm{Y}} = 43.8\%$).

This represents a huge compression of information into only a few summarizing score vectors, given that the original data contained approximately 300 000 observations. This is typical of our experience from analyzing large number of industrial batch processes using BWU. The significant information in the differences among the batches is often captured with only two or three latent variables. The corresponding analysis of $[\mathbf{Z}\ \mathbf{T}_{\mathrm{BWU}}\ \mathbf{Y}]$ is described in Section 2.10.3.2.4.

The superlevel block loadings[1] for the **Z** and **X** blocks show similar relationships to those in the loadings plot of **Figure 8**: The major contributors to deviations in quality captured by the PLS model come from deviations in the process variable trajectories and not from the incoming raw material variations. Clearly, the batch-to-batch variation in the operation of the process itself is the problem.

From any batchwise PCA or PLS analysis, each batch, i, is then uniquely summarized by a small set of score values ($t_{1i}, t_{2i}, \ldots$) that characterize the entire batch history. Comparison among batches can then be easily made by plotting their scores on joint score plots, as shown in **Figure 13** for the PLS of the drier data. In the score plots, any batches with approximately the same score values should have exhibited very similar time-varying covariance behavior over their batch histories. This allows for easy, direct comparison among the batches. For example, in **Figure 13** one can clearly see clusters: Batches 12 through 18 are very similar, batches 55 and 60 are similar, and batches 34 through 40 are similar. However, each of these clusters of batches behaved differently from one another.

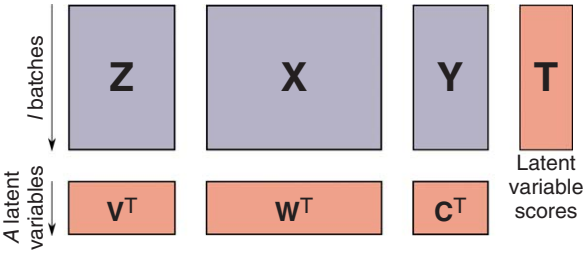

Figure 12 The general structure of batchwise unfolded (BWU) data. With the OWU–TBWU approach, the middle **X**-matrix comprises the BWU unfolded OWU-score matrix $\mathbf{T}_{\mathrm{OWU}}$.

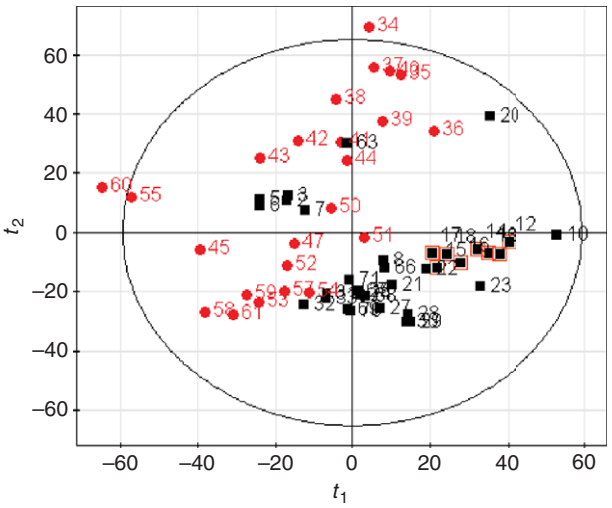

Figure 13 The score plot from the PLS model. The generally good separation between the on-specification (crosses) and off-specification (circles) batches is mainly along the t_1 direction.

More important is the clear separation visible in **Figure 13** among the good batches (to the right of the dividing line shown in the score plot) and the bad batches (to the left of the line). This mirrors the separation already shown in **Figure 3** obtained from a PCA on the final batch quality (**Y**) and shows that this separation in the quality data is clearly explainable from the batch trajectory deviations (**X**). (See **Figure 20** for very similar T_{BWU} results.)

From the score plot in **Figure 13**, it is apparent that good batches consistently have high values of the first score (t_1). The small cluster of good batches (2, 3, 5, 6, 7) at negative t_1 values was attributable to a specific operational variation unique to that cluster.[1] To interpret what process trajectory variations lead to good performance, one can look at the PLS loading weights (w_1^*) plotted in **Figure 14**. As can be seen from **Figure 12**, weights (w_1^*) are available for each variable at all time points during the batch.

In **Figure 14**, the w_1^* weights have been grouped by variable, and one can easily see the size of the weights for each variable and how the weights vary with time throughout the course of the batch. These time-varying weights reveal clearly the time varying, nonlinear nature of the information extracted from the batch trajectory data by the PLS model. It also enables one to easily see which variables have the greatest impact on the first score (t_1) and during which batch periods these variables have their effect. High values of t_1 (good batches) are

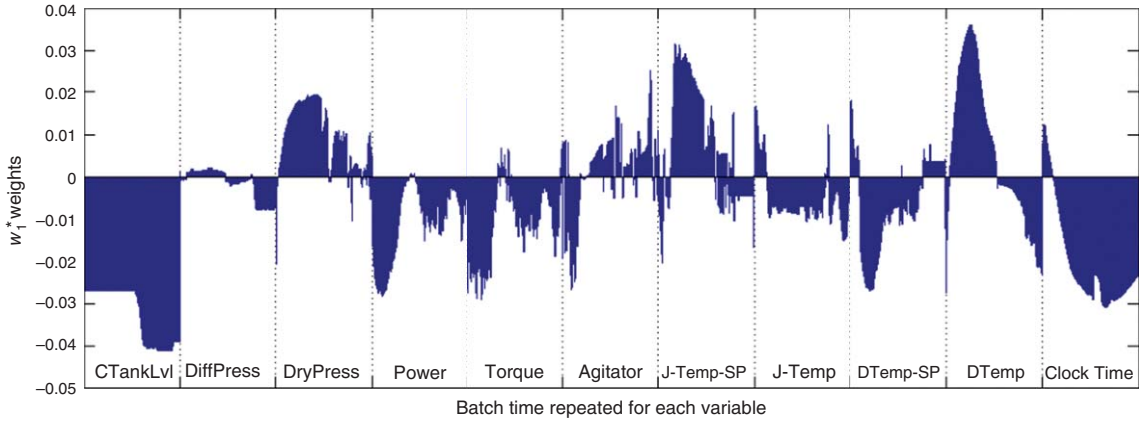

Figure 14 The w_1^* loading plot from the PLS model. High t_1 values correspond to good quality batch product.

clearly related to having low solvent collection tank levels (CTankLvl) throughout the entire batch (low w_1^* weights at every period) and high values of drier pressure (DryPress), jacket temperature set point (J-Temp-SP), and drier temperature (DTemp) during the first part of the batch (high w_1^* weights during this period) and low clock time (time usage) during the batch (low w_1^*). This allowed for diagnosis of the operating problems related to quality. For good quality product, it was important to have lower solvent amounts in the initial charge of wet cake (i.e., low solvent collected) and high levels of heating in the first phases of the process (leading to fast phase 1 completion times). If these conditions were adhered to, a good quality product was consistently achieved. More details on this example are found in Garcia-Muñoz *et al.*[1]

The three PLS models, one using landmarks, one using BWU trajectories, and one shown later using OWU followed by BWU of the scores, are in agreement. The score plots in **Figures 7**, **13**, and **20** show similar separation between the on-specification and off-specification product. Their corresponding loading plots also identify similar process changes that need to be applied to achieve good product quality.

Note that PLS analyses of both the BWU $\mathbf{X}$ and $\mathbf{T}_{\mathrm{OWU}}$ trajectories directly extracted three latent variables from the trajectories to summarize the differences among the batches, thereby avoiding the necessity of first prejudging and selecting landmark features from the trajectories (with the associated uncertainty that some important landmarks may be missed).

2.10.3.2.4 Contribution plots for BWU and TBWU models

As with steady-state processes, contribution plots play an important role in analyzing batch processes. However, for BWU the scores depend not just on the behavior of the process variable deviations at one instant of time but also on the time-varying behavior of the variable deviations over the entire batch duration (e.g., see w_1^* weight plot in **Figure 14**). This illustrates that batches can have different score values because of different behavior among the variables during any period of the batch. The contribution of any variable (x_j) to a difference in the t_1 score (Δt_1) between any two batches m and n is given by (in the $\mathbf{T}_{\mathrm{BWU}}$ models $\boldsymbol{t}_{\mathrm{BWU}}$ is used instead of $\boldsymbol{x}$).

$$\mathrm{contr}(x_j/\Delta t_{1;m,n}) = \sum_{k=1}^{K} w_{1,j,k}^{*}(x_{j;m} - x_{j;n}) \tag{1}$$

For the contribution to a simultaneous movement in more than one score direction, a generalization of this, weighting each score movement by its relative importance, is made (similar to that for steady-state processes). Now consider the two batches: batch 23 (on-specification batch) and batch 43 (off-specification batch). To investigate which variables are contributing most to the difference between them, we present the contribution plot in **Figure 15**. The difference between these batches is seen to be that, for the off-specification batch 43, the

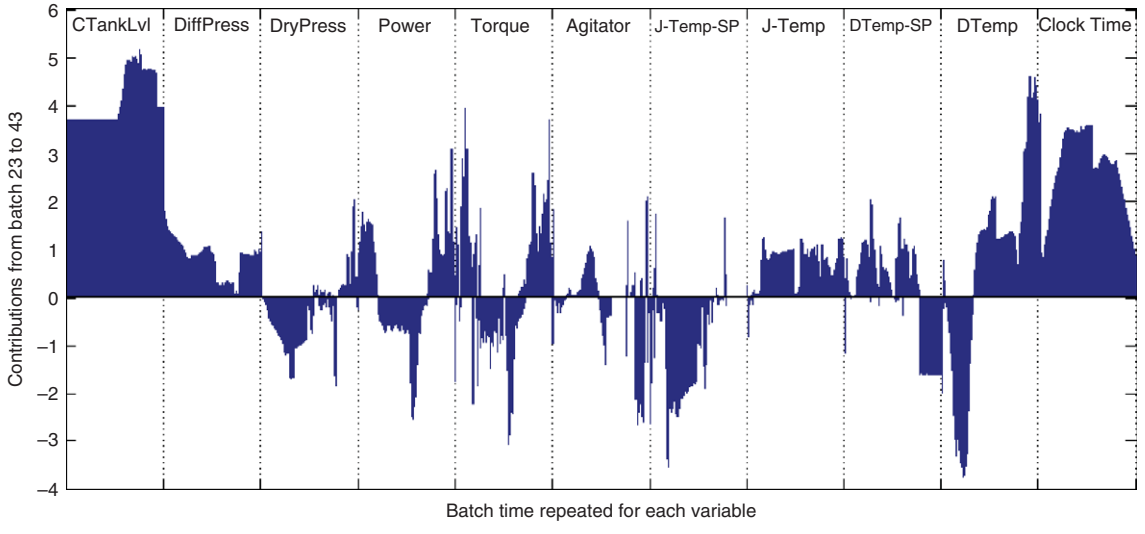

Figure 15 Contribution plot: from on-specification batch 23 to off-specification batch 43.

solvent collector tank level (CTankLvl) was much higher throughout the batch, the batch took much long time to complete (high clock time), and the heating was too low during the early stage and too high during the later stages (DTemp, J-Temp-SP). With T_{BWU} models, a further drill-down on the scores is needed to get the contribution values for the original variables over time (also see **Figures 17(a)** and **17(b)**).

2.10.3.2.5 OWU followed by BWU of the scores (OWU–TBWU)

This approach[9] first uses an observation level model to extract scores that summarize the process variables. It is particularly useful when online analytical sensors such as Mass or NIR spectrometers are used, and there are a resulting large number of spectral variables that can be compressed into a few latent variables.

Typically, PLS is used to develop a model of the scaled and centered OWU data (**Figure 9(b)**), using local batch time or batch maturity as the single **y**-variable.

The use of a **y**-variable related to batch completion and the PLS approach makes the first score describe mainly the linear behavior in **X**, that is, variables that go up *or* down monotonically will load heavily in $\mathbf{t}_1$. Analogously, variables with quadratic behavior, first increasing and then decreasing or *vice versa*, will load heavily in the second score. Cubic behavior with time will be captured in the third score, and so forth. Hence, the use of PLS with a relevant **y** leads to a certain mildly imposed structure on the resulting scores, which would not result from a PCA. This also makes the score parts of each batch comparable.

In a second stage, a second multivariate analysis is made of the BWU scores (recentered and rescaled) *or* the BWU raw data (centered and scaled) or a combination of these. If the batches are of different length, the OWU scores (or $\mathbf{X}_{BWU}$) first need to be aligned (Section 2.10.3.2.2). When the data are divided into phases, the unfolding and alignment of the scores is done separately for each phase.

This gives a second model (batch-level model) showing similarities and dissimilarities between the completed batches with all the data as described above in Section 2.10.3.2.2. Here PCA or PLS is used depending on the absence or presence of quantitative variables, **Y**, describing the final quality of the batch. When the batch data are divided into phases, this second batch-level analysis is often done hierarchically,[4] with one or several blocks per phase followed by a final BWU model using the block-scores as variables.

The purpose of these two analyses and the resulting two models is to (1) understand the evolution of each batch, and simultaneously compress a multidimensional **X** into a few columns score matrix **T** (OWU analysis), and (2) compare the whole batches for final quality assessment, troubleshooting, historical analysis, and so forth, like in Section 2.10.3.2.2 above (BWU). Also, step (1) provides a model that is useful for online batch SPC, discussed in Section 2.10.5.2.

The OWU–TBWU approach is illustrated by the yeast example, but examples 1 and 2 were also analyzed by this approach to provide comparisons with the other approaches.

2.10.3.2.5(i) Capturing nonlinearities and changing covariance structures (dynamics) in OWU models using phases To handle the changes in the covariance structure of **X** during the batch evolution, as well as other nonlinearities in OWU models, it is practical to divide the batch evolution into phases and develop a separate model for each phase. If a sufficient number of phases are used, the changing covariance structure and other nonlinearities can be well approximated. This includes separate centering and scaling for each phase and possibly also different data transformations, as well as separate alignments of OWU scores or BWU raw data for the second upper batch-level model. Except for alignment, the BWU models need no division into phases. See example 1 in Section 2.10.3.2.2.

In this way, the OWU–TBWU modeling needs more care on the lower level than the single model BWU approach. Indeed, when the number of process variables, $\mathcal{J}$, is small, it is often preferable to use the unfolded **X** in the upper batch-level modeling instead of the unfolded score matrix, $\mathbf{T}_{OWU}$.

In early applications of batch data analysis with the OWU approach, this problem with changing covariance structure and the need for division of the data into phases was often overlooked. However, with the inclusion of the phase option in the OWU approach, the selection of an adequate number of phases and the selection of an adequate number of components in each phase, the OWU–TBWU and the BWU approaches give equivalent result both in modeling similarities and dissimilarities between batches, and in the online monitoring of the batch evolution. The comparisons between the two approaches made on the three examples are summarized in Section 2.10.3.2.5(vi), and indeed, there is no practical difference between them.

2.10.3.2.5(ii) Separate OWU analyses of different kinds of data When the evolving batches are characterized by data of widely different types and formats, for example, 10 process variables and 3000 NIR spectral variables, it is essential to make a separate OWU analysis of each type of data, giving a score matrix **T** for each type, which then is aligned and unfolded to give BWU scores of several types. If separate OWU-PLS analyses are not made, great difficulties ensue in the modeling and interpretation on both the lower OWU level and the upper BWU level.

This is analogous to the situation with spectroscopic data in the initial BWU approach (Section 2.10.3.2.2). There an initial preprocessing by means of a PCA or PLS of the spectral data is needed, so as not to swamp the upper level model and mask the effects of the much fewer process variables.

Also, as described below, when the process variables are fairly independent or few – say, fewer than 15 – it is common to use the original process variables on the upper batch level in the same way as the BWU approach above (Section 2.10.3.2.2).

2.10.3.2.5(iii) Miscellaneous points With controlled process data where the set points are changed as part of the process control scheme, it may be advantageous in the centering of the data to subtract the set-point values for each batch instead of subtracting the total averages from **X**. This partly removes the **X**-variation caused by the set-point modifications and makes the batches more comparable.

In the same context, one may get a better idea of the acceptable ranges for controlled process variables by using a fraction of their set-point values, say 5 or 10%, as the basis for their scaling rather than their standard deviations.

The number of PLS components, A, in the OWU stage is determined by a combination of CV related to the prediction of **y** by **X**, and the side condition that the variation of **X** should be modeled to at least, say, 75–80% (often more) with a limited number of components. This is because the resulting scores, $\mathbf{t}_a$, should capture sufficiently much of **X** for the second batch-level analysis (see Section 2.10.3.2.5). In case the **X**-variation is not adequately captured by the PLS model, we recommend that the upper level batch model is based on the original **X**-variables, and hence, this second model becomes the same as the one in Section 2.10.3.2.2.

2.10.3.2.5(iv) The baker's yeast example The three-way matrix was OWU (**Figure 9(b)**) giving a training set having 7 variables and 1660 observations (20 batches × 83 time points) in the unfolded matrix $\mathbf{X}_{OWU}$. An additional matrix with 1079 rows (13 batches × 83 time points) and 7 columns constitutes the prediction set. The data were not divided into phases.

The $\mathbf{X}_{OWU}$-data matrix is augmented by a single column **y**-matrix, preferably containing the values of a maturity variable such as yeast cell count or cell density. In the absence of a maturity variable, the local batch time is used as a maturity variable surrogate. This restarts at zero at the beginning of each batch and provides the time after start in arbitrary units for each observation data vector (with $\mathcal{J}$ elements). This was done in the yeast example.

X-variables with large variation may be logarithmically transformed before the analysis, as done with ethanol and NH_3 in the yeast data. Thereafter, the training set was preprocessed by centering and scaling all variables by subtracting their averages and dividing by their standard deviations. The prediction set ($N_{batch} = 13$, $\mathcal{J} = 83$, $N2_{obs} = 1079$) was transformed, centered, and scaled using the training set recipe of log transforms, averages, and standard deviations. The preprocessed training data are subjected to an ordinary PLS analysis with the selected **y** variable (local time).

With the yeast example, $A = 3$ significant (according to CV) components result, where 79% of the **X**-variance is explained. With seven **X**-variables, this is a contraction to approximately 40% the original raw data in the later analysis of the unfolded scores. Hence, the use of three scores to represent the batch dynamics corresponds to a substantial filtering. Six of the **X**-variables were modeled to $R^2 = 0.7$ or more, while pH had $R^2 = 0.57$ because of a limited variation. No nonlinearities were seen, and hence, no division into phases was done.

For the interpretation of the OWU-PLS model, the resulting scores and loadings can be plotted in bivariate or 3D scatter plots. Plots of the scores against time show the evolution of the batches. As the main purpose with the modeling of the yeast data was process monitoring, these plots are shown in Section 2.10.5.2.

Before the second batch-level analysis, the OWU scores must be (1) aligned (not needed with the yeast data) and then (2) folded and BWU. With three OWU scores, each batch is summarized as 3 × 83 values. Laying these out as 3 × 83 **X**-variables per batch, and augmenting these with (1) additional **Z**s being the initial variable values for each batch – here the single inoculate value – and (2) response values (**Y**s) – here yield – gives one

row-vector per batch, divided into **X**s and **Y**s (responses), just like **Figure 12** but with the OWU-T instead of the BWU **X** in the middle. In the yeast example, this gives 250 variables (3 scores × 83 time points plus the single starting inoculate value) for the **X**-matrix and a single **y**-variable (yield). Below, we describe the analysis of these data by PCA (for classification of batches as acceptable or not) and PLS (for relating the process data, **X**, to **y** = yield). Before the analysis, the data are centered and scaled as usual.

The PCA of the OWU–TBWU – part of the training set ($K = 250$) – gave $A = 3$ significant components (CV), describing $R^2 = 73\%$ and predicting $Q^2 = 52\%$. Approximately half of the individual variables were well explained with R^2 and Q^2 values exceeding 0.8 and 0.6, respectively. Less well explained are time points 35–50 of t_1, and the first and last 15 time points of t_3.

The classification of the 13 prediction set batches by their predicted residual standard deviation, DModX, is displayed in **Figure 16**. The DModX does classify all prediction set batches correctly, but one of them

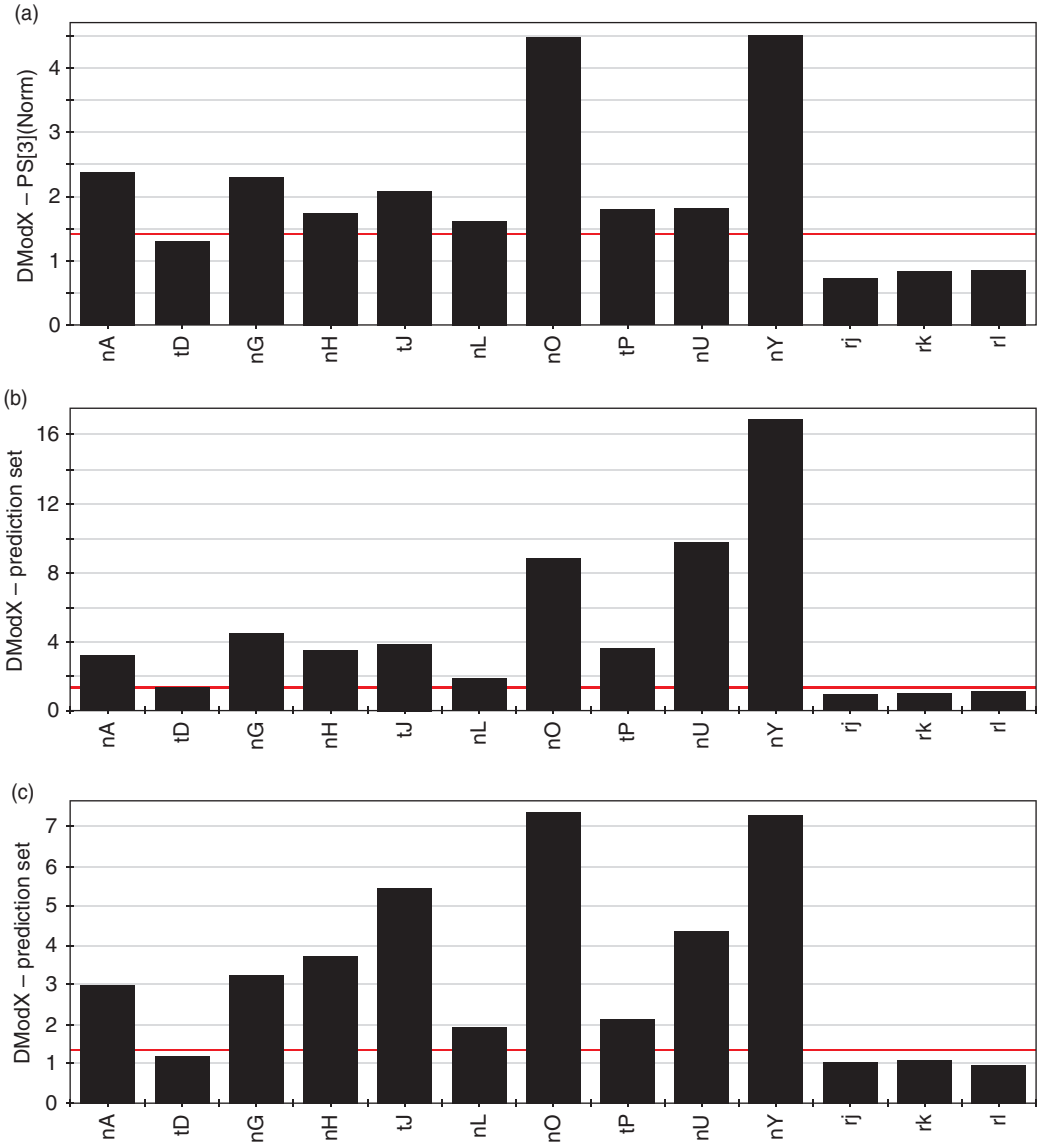

Figure 16 Resulting classification based on the DModX (SPE) of the 13 prediction batches of the yeast example by the three batch-level models: (a) landmark approach, (b) BWU of **X**, and (c) OWU–TBWU. The red lines indicate the 95% probability levels.

(the good D) is too close to the decision limits to provide a certain classification. The use of the Hotelling's T^2 based on the predicted score values does not improve the classification and is hence not shown.

Figure 16 also includes the results of a landmark feature-based (Section 2.10.3.1) and the BWU approaches, both using PCA on the same training and prediction sets. The results are similar, but the differences between acceptable and nonacceptable batches are smaller for the landmark approach.

The interpretation of individual problem batches is greatly facilitated by the contribution plot, where the scaled **X**-profile of the specific batch is compared with a reference batch, usually the average good batch. This contribution plot of batch **Y** (**Figure 17**) indicates deviations in t_2 culminating at time 52. These in turn when drilled down into the OWU PC model (lower plot) indicate problems with the ethanol and molasses.

Second, a PLS analysis was made using the same data as **X**, plus the y-variable, yield. Three training set batches had no value of yield and were hence excluded. A reasonable one-component model is obtained, with R_y^2 and Q_y^2 being 0.52 and 0.11, respectively. **Figure 18** shows the PLS-regression coefficients for this model. Coefficients for unfolded scores from the start, middle, and the end of the batch evolution dominate. Using the lower level loadings (**Figure 18**) to calculate the coefficients for the individual variables at points 6, 38, and 75 shows ethanol dominating the start and the end, while molasses feed and NH_3, all with positive coefficients, dominate the middle.

The plots of y observed versus predicted for the training and prediction sets are shown in **Figures 19(a)** and **19(b)**. The prediction set error (RMSEP = 0.028) is somewhat larger than the one of the

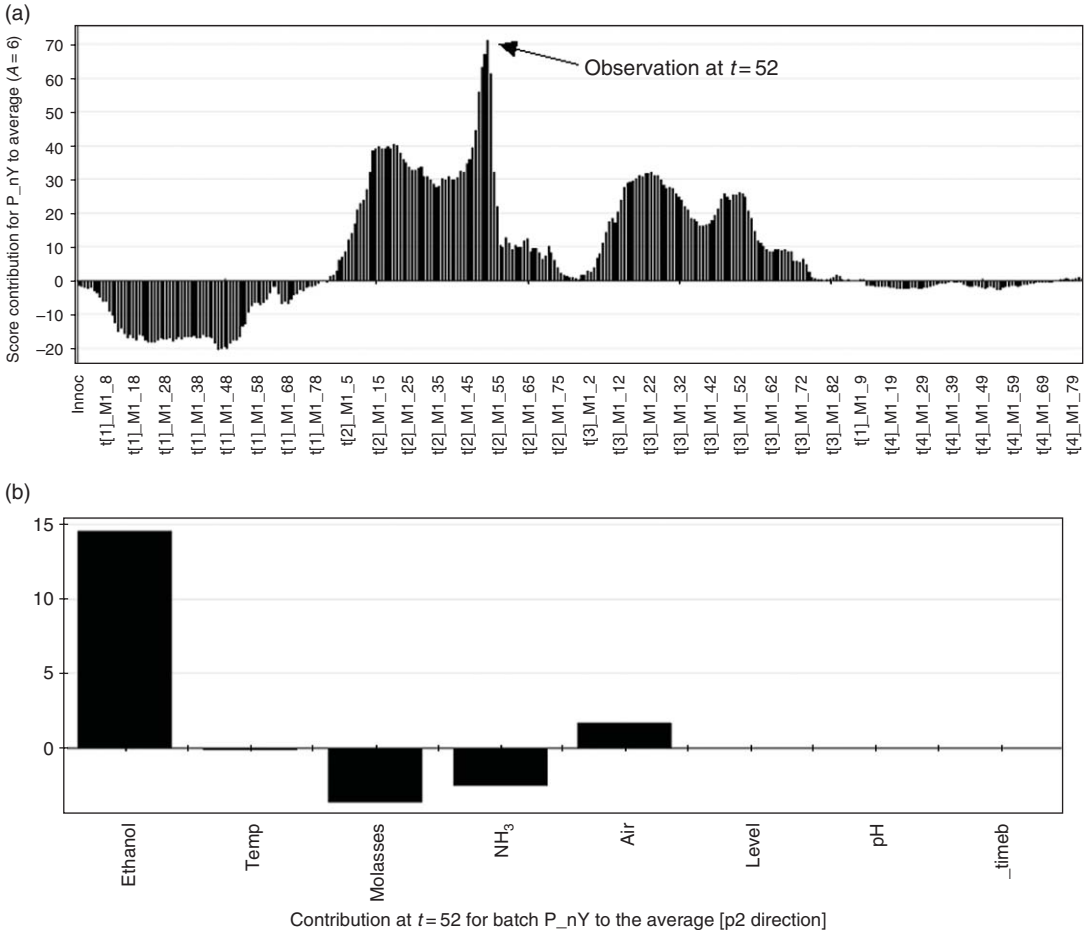

Figure 17 (a) Contribution plot calculated by the batch level QWU–TBWU PC model for yeast batch Y related to the average good batch. (b) The maximum at time 52 is resolved into the corresponding OWU-level contribution plot.

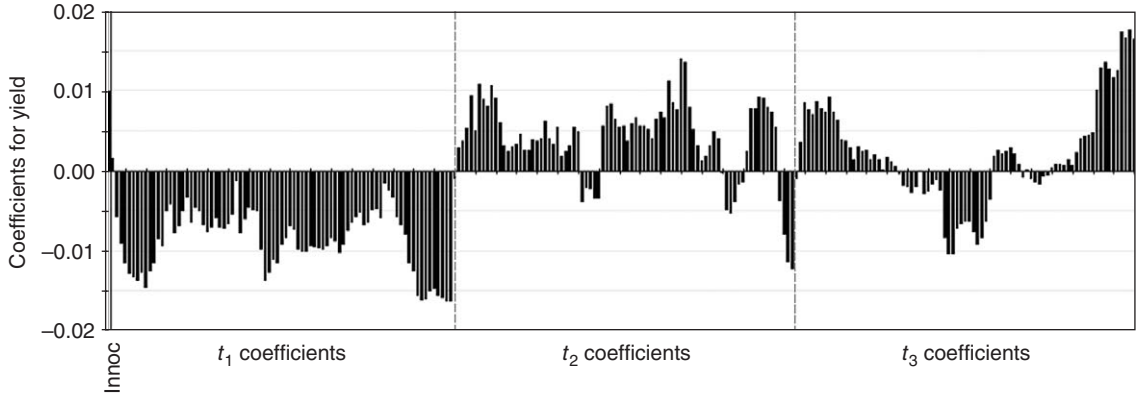

Figure 18 PLS regression coefficients of the upper level batch model.

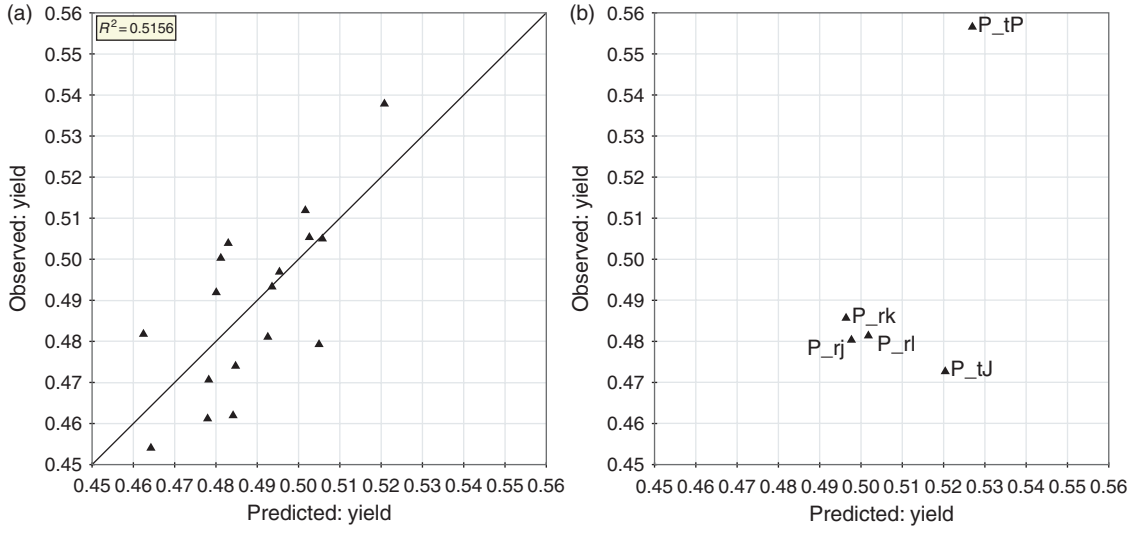

Figure 19 y-Observed against y-predicted for (a) the training data and (b) the prediction data.

training set (RMSEE = 0.015). However, the presence of two bad batches (*J* and *P*) in the test set makes the judgment of these results difficult.

2.10.3.2.5(v) OWU–TBWU analysis of the drier data (example 1) The first observation level analysis of the OWU **X** data of example 1 with $K = 10$ **X**-variables was done by using three PLS models, one for each phase. Each gave a five-component model (significant by CV), with R^2_X values of 0.743, 0.748, and 0.686, respectively. All variables except two were explained to more than 50%, most to more than 80%. The two exceptions were DiffPress in phase 1 and J-Temp in phase 3, where both were almost constant.

The second batch-level analysis based on unfolding the three OWU score matrices (three phases), followed by a PLS analysis of **Z**, $\mathbf{T}_{\text{BWU}}$, and **Y** gave three components, significant according to CV. The plot of the first two PLS scores (**Figure 20**) shows very much the same separation of good and bad batches as previous analyses of the landmark and the BWU **X** data (**Figures 7** and **13**), with the only difference that the first score has the opposite sign. These are summarized in **Table 3**.

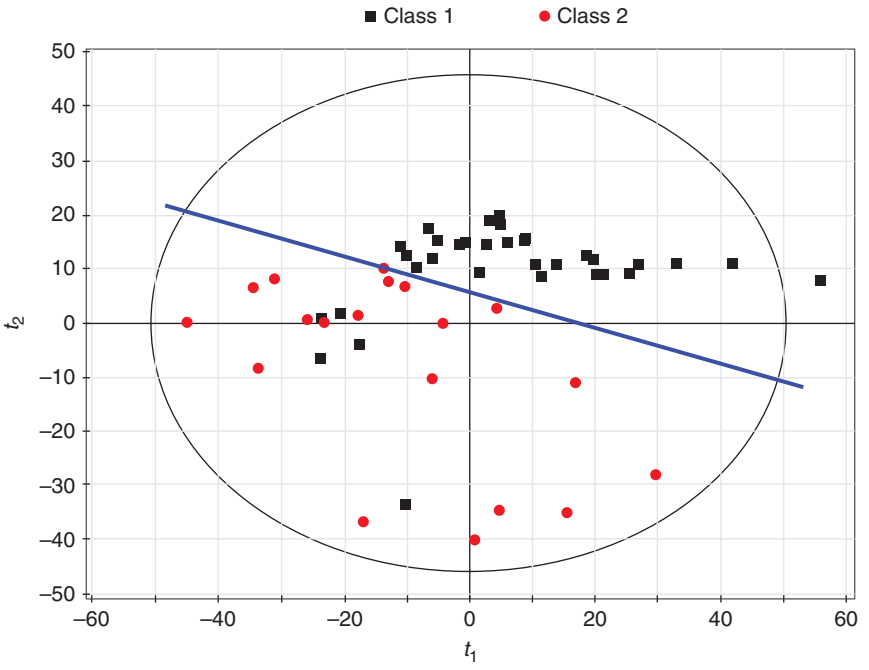

Figure 20 The two first PLS scores of the upper batch model where the unfolded score matrix **T** from the lower level comprises the **X** matrix together with the initial conditions, **Z**. Compare with **Figures 7** and **13**.

Table 3 Summary of results for the drier data set, batch level models (example 1), and yeast data set, batch level models (example 3)

Approach	*A (number of PLS components)*	*R^2X (explained X variance)*	*R^2Y (explained Y variance)*	*Q^2Y (cross-validated R^2Y)*	*RMSEP, test set*
Drier data set					
Landmark	1	0.250	0.502	0.092	0.028
	2	0.389	0.801	0.369	0.030
BWU of **X**	1	0.260	0.431	0.038	0.023
OWU–TBWU	1 (3 phases)	0.187	0.516	0.11	0.028
Yeast data set					
Landmark	1	0.250	0.502	0.092	0.028
	2	0.389	0.801	0.369	0.030
BWU of **X**	1	0.260	0.431	0.038	0.023
OWU–TBWU	1	0.187	0.516	0.11	0.028

2.10.3.2.5(vi) Comparisons between the landmark, BWU, and OWU–TBWU approaches **Table 3** shows a summary of the results of the three approaches for examples 1 and 3 (Drier and Yeast). The nylon example was not included because of the low precision of the **Y**-data.

As described above, the data were analyzed in the following way:

1. Landmark analysis (examples 1 and 3);
2. BWU of the batch-data matrix **X** followed by centering and scaling and PCA (example 3) or/and PLS (example 1 and example 3);
3. OWU unfolding of **X** (three phases in example 1), followed by PLS using $y =$ maturity (phases 1 and 2 in example 1) or local time (example 3), followed by BWU of the resulting score matrix or matrices, and using the unfolded scores as variables in an upper level batch model using PCA (example 3) or PLS (examples 1 and 3).

We should note that the CV results of the OWU–TBWU approach are not comparable with the BWU and landmark approaches because in the former only the second step is included in the CV of the second batch-level model. Hence, a second comparison has been made for examples 1 and 2 with two independent models and two independent prediction sets.

The results for the batch-level models are summarized in **Table 3**. In the first example (drier process), it is seen that the BWU and the OWU–TBWU models each has three components, and the results are similar.

Example 3 shows diverse results, with the landmark model looking the best from the training set results, while the BWU model looks the best with the test (prediction) set. The presence of two bad batches ($\mathcal{J}$ and P) in the test set makes the interpretation difficult.

Overall, to provide an unambiguous comparison between the two unfolding approaches, the data of example 1 (preexcluding the 12 outliers) were divided into two parts, one with odd and the other with even indices, respectively. Two analyses were then made for each approach (BWU and OWU–TBWU) in the same way as described earlier (three phases for the OWU model) using the odd as the training set and the even as the prediction set, and *vice versa*.

The two methods give closely similar results in the predictions, with no method being significantly better. For the eight responses, each approach was slightly better than the other 8 of the 16 times (two models with eight **Y**s). The average difference between the 2 over these 16 times was 1.7% (slightly lower prediction errors for the OWU–TBWU approach).

So, in summary, these comparisons confirm the close similarity between the two BWU approaches.

2.10.3.2.6 Advantages of BWU for analysis

The advantages of BWU of either the raw data **X** or the OWU scores **T** for analyzing historical batch data are now apparent. The resulting latent variable models capture the time-varying nonlinear dynamic behavior of the batch process; they focus on covariance differences among the batch histories, thereby allowing for comparison among batches; they allow for the inclusion of initial conditions (**Z**) and final quality properties (**Y**) for each batch (row of the matrices) and hence for PLS modeling.

2.10.3.2.7 Other three-way methods

In this chapter, we have described in some detail the more commonly used methods for dealing with the batch three-way array **X** based on unfolding it into two-way matrices which are thereafter analyzed using PCA and/or PLS. Other direct three-way latent variable methods such as PARAFAC and TUCKER3 have occasionally been used for modeling the three-way **X** array data from batch processes.

However, these approaches (especially the restrictive PARAFAC model) are generally not well suited for analyzing batch processes. They impose certain restrictions in the model that will rarely be satisfied in batch processes. In particular, PARAFAC assumes that, in each dimension, the relative loadings among the variables are constant throughout the entire batch duration (i.e., that the batch covariance structure is time invariant). However, as with OWU approach (which makes the same assumption for its first lower level model), the batch could be broken up into phases each having a more constant covariance structure among the variables, and then multiple models built, one for each phase.

We shall not further pursue direct three-way methods, but refer to Westerhuis *et al.*,[13] who compared the BWU-PCA/PLS approaches with the PARAFAC and TUCKER3 approaches in some detail.

2.10.4 Online Prediction of Future Trajectories and Final Quality

In many situations during the batch, one is interested in predicting the future behavior of the batch trajectories and the final product quality. In particular, these are of interest in process monitoring and process control (see Sections 2.10.5 and 2.10.6). In this section, we present the basic ideas and methods for making these predictions.

The objective is to provide predictions during the course of a new batch. Hence, latent variable models with BWU are required, that is, either the BWU or the OWU–TBWU approaches. These model the dynamic, time-varying trajectories and the final quality for each batch. We will here consider only PLS models, although if

there is only an interest in predicting future process variable trajectories, then BWU PCA models can be used. However, the method is the same. We assume here that a BWU PLS model using $\mathbf{Z}$, $\mathbf{X}$ or $\mathbf{T}_{\mathrm{OWU}}$, and $\mathbf{Y}$ data has already been built from a representative data set of historical batches.

The main issue in the prediction of final batch quality using any BWU approach, when currently at time point k during a new batch, is that the trajectory data on all future variables from time $k+1$ up through the final batch time K are missing. The scores (t_1, t_2, ...) and final product quality ($\mathbf{Y}$) for any batch from the BWU PLS model are functions of all the variables over the complete batch history. Therefore, the batch model should be used to impute (predict) the remaining trajectories of all the variables, and then, from these imputed values, the scores and final quality prediction can be computed.

There are numerous approaches for imputing the missing variable trajectories over the remaining period on the batch. Aside from a few ad hoc approaches that make rather strong assumptions (such as assuming that the future trajectories will just be the mean trajectories, filling in with zeroes, or maintaining the current deviations for the remainder of the batch), the methods are all based on missing data imputation methods for PCA and PLS models.[14,15] Nomikos and MacGregor[6–8] used and recommended the projection to the model plane (PMP) approach in their early work, but recent research has shown that conditional mean replacement (CMR) or trimmed score regression (TSR) approaches may be better[15] in certain cases. We present and illustrate just the PMP approach in this chapter, and so the readers are referred to the literature for other algorithms.

Suppose, during a new batch we are currently at time k and have observed the trajectories on the $\mathcal{J}$ process variables up to that time (i.e., $\mathbf{x}^{\mathrm{T}}_{\mathrm{new},1:k\mathcal{J}}$). The objective is to impute the missing data trajectories for the remainder of the batch assuming them to be consistent with the already observed values up to time k ($\mathbf{x}^{\mathrm{T}}_{\mathrm{new},1:k\mathcal{J}}$) and the covariance structure of the PCA or PLS model for the remainder of the batch. By partitioning the complete row vector $\mathbf{x}^{\mathrm{T}}_{\mathrm{new}} = (\mathbf{x}^{\mathrm{T}}_{\mathrm{new},1:k\mathcal{J}}\ \mathbf{x}^{\mathrm{T}}_{\mathrm{new},(k+1)\mathcal{J}:K\mathcal{J}})$ into its known and unknown parts and correspondingly partitioning the loading and weight matrices of PCA/PLS, it can be shown[16] that the least squares estimate of the $1 \times A$ score vector row $\tau^{\mathrm{T}}_{\mathrm{new}}$ for the new batch, based on information up to time k, using PMP, is given by

$$\hat{\tau}_{\mathrm{new},k} = \left(\mathbf{P}^{\mathrm{T}}_{1:k\mathcal{J}}\mathbf{P}_{1:k\mathcal{J}}\right)^{-1}\mathbf{P}^{\mathrm{T}}_{1:k\mathcal{J}}\mathbf{x}_{\mathrm{new},1:k\mathcal{J}} \tag{2}$$

The time-varying covariance structure of the complete batch can be partitioned as

$$\mathbf{x}^{\mathrm{T}}_{\mathrm{new}} = \left[\mathbf{x}^{\mathrm{T}}_{\mathrm{new},1:k\mathcal{J}}; \mathbf{x}^{\mathrm{T}}_{\mathrm{new},(k+1)\mathcal{J}:K\mathcal{J}}\right] = \tau^{\mathrm{T}}_{\mathrm{new}}\left[\mathbf{P}^{\mathrm{T}}_{1:k\mathcal{J}}; \mathbf{P}^{\mathrm{T}}_{(k+1)\mathcal{J}:K\mathcal{J}}\right] \tag{3}$$

and an estimate of the remaining trajectory data that is consistent with the model and the history of the current batch ($\mathbf{x}^{\mathrm{T}}_{\mathrm{new},1:k\mathcal{J}}$) is given by Garcia-Muñoz *et al.*:[16]

$$\hat{\mathbf{x}}^{\mathrm{T}}_{\mathrm{new},(k+1)\mathcal{J}:K\mathcal{J}} = \hat{\tau}^{\mathrm{T}}_{\mathrm{new}}\mathbf{P}_{(k+1)\mathcal{J}:K\mathcal{J}}{}^{\mathrm{T}} \tag{4}$$

These imputed future batch trajectories are very different from simple linear multivariate time series predictions based on past data.[16] They incorporate all the past data up to time k, but in addition they utilize the known time-varying weight and loading matrices over the future horizon up to the end of the batch to make the future predictions. In that sense, the imputations are more of an interpolation than an extrapolation, and they are highly adaptive as new data become available.

The nylon batch polymerization data are used here to illustrate the accurate, nonlinear, time-varying future trajectory predictions that can be obtained. A PCA model on the BWU matrix of the 10 variables was built with all batches, except for some outlier batches (37, 39, 43–55) and the batch used to illustrate the trajectory predictions, batch 22. Given the data available on all variables up to time 30, **Figure 21** shows the predictions made for the remainder of the unmodeled batch for ($k = 31, \ldots, 100$) on one of the variables. Similar predictions are available for the other trajectories. The one plot shows the complete trajectory predictions with the average trajectory added to the deviation predictions, and the other plot shows the future predictions of just the variable deviations from the mean. It is evident that the early behavior of the measured trajectories, and the time-varying loadings are being used for providing very accurate predictions of the future deviations in the trajectories.

The prediction, made at time k, for the final product quality vector for the new batch (at time K) is then given by:

$$\hat{y}^{\mathrm{T}} = \tau^{\mathrm{T}}_{\mathrm{new}}C \tag{5}$$

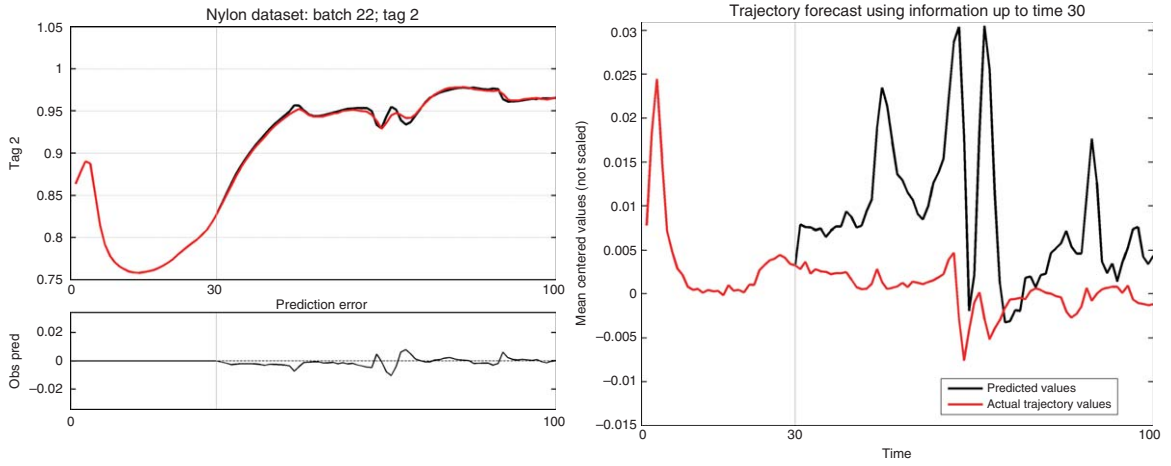

Figure 21 (a) The trajectory predictions from $t = 30$ onward in the original units of the variable. (b) The predictions of the variable from $t = 30$ onward, but shown as deviations from the mean trajectory.

This enables one to provide updated online predictions of final product quality continually as the batch progresses.[8] A similar approach can be used for predicting the final end point (harvest time, etc.) to achieve the best quality.[17] These predictions will be a critical part of the batchwise monitoring, control, and optimization problems discussed in the following sections.

An alternative approach is to build separate BWU (or TBWU) PLS models for different degrees of completion of the batch (e.g., 50, 75, and 90%) and then predict the final product quality when each of these models has complete data (e.g., at the 50, 75, and 90% completion points). However, this stagewise fixed model approach is rarely as good as using the full BWU model with missing data imputation. The stagewise approach does not use knowledge of the changing loading structure for the remainder of the batch and hence cannot predict as well how disturbances evident in the early part of the batch will propagate through the remainder of the batch. On the contrary, the full BWU model has this knowledge, provided it is based on representative data.

2.10.5 Monitoring (MSPC) of Batch Processes

After the analysis of historical batches has been completed and all improvements to the process and operation have been made, it is desirable to set up a MSPC monitoring scheme to maintain these improvements and to detect and diagnose any new abnormal behavior in the future. To accomplish this, one needs a multivariate model for the common cause behavior of the process under normal operation. The models that were used for process analysis and troubleshooting in Section 2.10.3 are therefore not appropriate. They were based on historical data from both normal and abnormal batches before improvements were made. To establish a multivariate SPC monitoring scheme, PCA or PLS models based on new batches where only common-cause variation is present must be used. Control limits in the score space, Hotelling's T^2, and on the SPE or DModX can then be used to assess whether future batches perform in a manner consistent with normal (common-cause) variation, contribution plots, and so forth, can be used to help diagnose the possible cause for any detected abnormality.

There are basically two approaches for monitoring batches. One is performed only at the end of each batch to check whether it was in a state of statistical control. If any abnormal behavior is detected, then analysis of the batch can be performed immediately to try to diagnose the cause of the problem and correct it for the next batches. Furthermore, the results can be used to segregate the batch product to storage or to subsequent usage based on the results. This approach is typical for fast processes, for example, in semiconductor manufacturing, where a batch takes only a couple of seconds, and there is no time to react during the batch evolution,[18] and for

processes in which there is no mechanism for online control or modification of a poorly performing batch after it is detected (see Section 2.10.6 for online control).

The other approach is to perform online monitoring in real-time at all or at several selected time points during the progress of every new batch. In this way, the progress of each new batch can be followed, and abnormalities detected and potentially diagnosed while the batch is still running. This additional real-time monitoring is advantageous if one has the ability to act on the knowledge gained during the progress of the batch, either through online control (see Section 2.10.6) or some other form of action made before the end of the batch (e.g., early termination).

For the first option (monitoring only at the end of each batch) any of the above analysis methods described in Section 2.10.3 can be applied in a straightforward manner to evaluate the batch since the new batch is complete. On the contrary, for online monitoring, several other factors come into play. The first is that a real-time data collection system must be in place. Another is that the method used must accommodate for the fact that at any time point during the batch, the data for the remainder of the batch will be missing. Therefore, the BWU method will have to incorporate some form of missing data imputation to obtain the scores, T^2, and contribution plots (Section 2.10.4).

In this chapter, we focus only on the real-time monitoring applications, as offline (end of batch) monitoring is straightforward.

2.10.5.1 Monitoring Based on BWU

BWU methods include the landmark approach and any of the BWU multiway PCA/PLS approaches described in Section 2.10.3.

2.10.5.1.1 Landmark approach

The landmark approach can be easily used in a modified form for online monitoring. The model with landmarks is built from batches with only common-cause variations. For a new batch, monitoring can be done after any time point where new landmark data become available. The remaining landmarks for the batch are simply treated as missing data (see Section 2.10.4) and the scores and DModX values updated at each point and compared with their control limits. These limits can be computed in a manner similar to those described below in the BWU PCA/PLS approach.

2.10.5.1.2 BWU-based PCA/PLS models

The monitoring approach based on BWU **X** and PCA or PLS models was developed by Nomikos and MacGregor.[6–8,19] Again, a BWU model is built using only batches that exhibit common-cause variation. This model can then be used to assess new batches either as an offline method performed upon the completion of each new batch or as an online monitoring scheme to track the performance of the new batch at every time point during its evolution.

For the former approach, the batch-to-batch assessment is performed just as described in the analysis section for complete batches, namely by assessing the scores (score plots or Hotelling's T^2), and the SPE or DModX for that batch against the corresponding control limits, and diagnosing any detected deviations using contribution plots.

For the latter (online) approach, the additional step of imputing the process variable trajectories for the remaining portion of the batch is required at each time step (see Section 2.10.4). The imputed score values and DModX are then compared against their time-varying control limits for that period. BWU automatically captures the time-varying covariance structure among the variables, and historical batch data can be used to define appropriate time-varying confidence limits on the score and SPE charts. An approach for developing these control limits based on the use of historical, in-control batch data is described by Nomikos and MacGregor.[8] In this approach, all these in-control historical batches are passed through the BWU monitoring scheme, and control limits are defined that include a selected percentage of these good batches (e.g., 99%).

We now illustrate the use of this BWU approach on the batch nylon polymerization data. A BWU PCA model ($A = 3$ components) was developed from data on 36 historical batches (batches 1 through 36) that represented common-cause variation and provided good final product. The monitoring of a new batch, batch

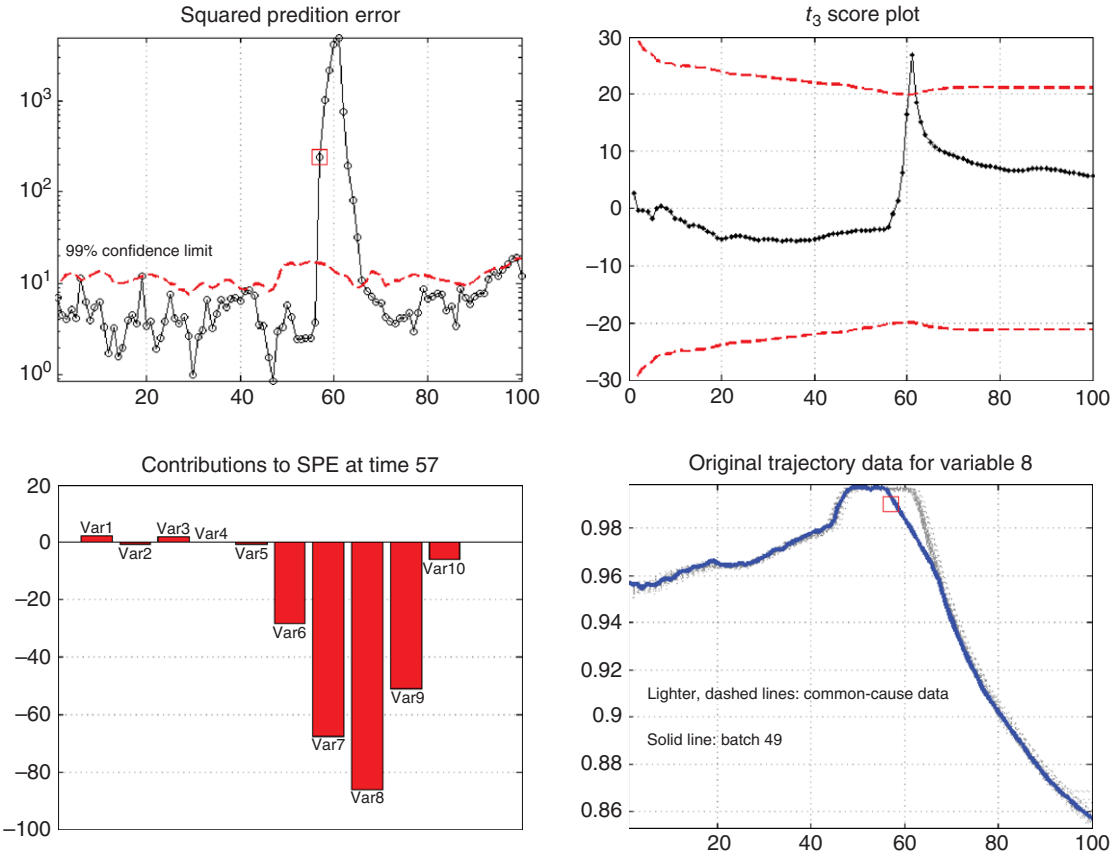

Figure 22 Online, real-time process monitoring for batch 49. The SPE and *t*-scores are monitored throughout the batch period. Contribution plots are drawn as soon as a limit violation occurs at time 57 (square point). The pressure variable 8 has the strongest contribution and the original time-series data confirm the problem. Other variables 6, 7, and 9 are highly correlated with variable 8, and the problem related to this group of variables is that they are too low during this period.

49, shows a subtle variation in the trajectories of variables 7 through 9. This batch yielded a product with barely acceptable material. The monitoring plots in **Figure 22** show that this batch came back into the control limits after the deviation; however, that does not imply the batch product will be good. The deviation between times 57 and 66 clearly could have a significant effect on the progress of the batch and on the final product.

Abnormal behavior in the batch is detected around time 57 in both the SPE and the score plots. A contribution plot for the deviation in the SPE at time 57 is shown in **Figure 22**. This clearly shows that low values of the variables 6, through 9, starting at time 57 and continuing through to time 65, are the main variables related to the abnormal event.

2.10.5.2 Monitoring Based on OWU

On the lower observation level of an OWU model, the OWU scores, residuals, Hotelling's T^2, and DModX (SPE) are immediately available when the data of a new observation vector have been collected. It is therefore straightforward to construct SPC limits at any degree of completion of a batch. The only difficulty arises when batches differ much in duration, as the scores cannot be aligned until the completion of the batch (or phase). Hence, the SPC charts on the lower level always show the results unaligned.

If a prediction batch goes outside the SPC limits part of the time, the integrated part outside the limits relative to its area inside the limits provides a useful summary diagnostic called OOC for out of control. This is shown in **Figure 23** (right) for the yeast data prediction set Hotelling's T^2 and DModX.

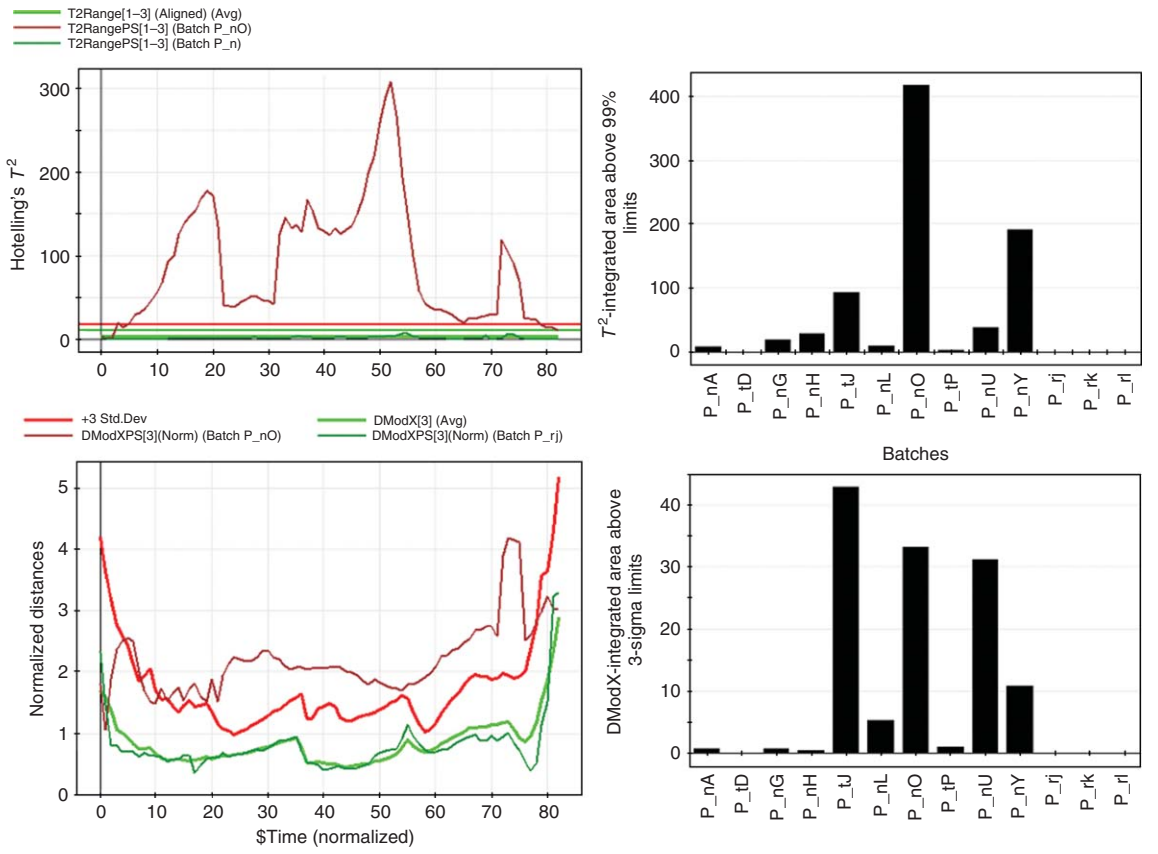

Figure 23 Left: OOC Hotelling's T^2 and DModX for the two prediction-set batches j and O. The upper two plots are Hotelling's T^2 and the lower two plots are DModX. The green curves show the average trace over the training set batches whereas the red lines indicate the 3 sigma limits. Right: The integrated area above the limits for all 13 prediction-set batches.

The combination of the OOC Hotelling's T^2 and DModX plots are seen to provide a good quality diagnostics at the lower observation level for these yeast batches. For the complete batch, they form a complement to the diagnostics based on upper level batch models (Section 2.10.3.2.6).

2.10.5.2.1 Score plots and loading plots for the OWU PLS model

The score plots versus time summarize the evolution of one or several batches. These can be shown as aligned (see Section 2.10.3.2.2) or unaligned values. For the aligned vectors, their average and standard deviation at each time point is calculated. SPC limits based on the average batch (e.g., ±3 standard deviations) at each time point are displayed in the plot. The training set of acceptable batches should usually fall inside the SPC limits at all times. **Figure 24** shows an example of a good batch (black). We note that the first and the second scores show approximately linear and quadratic trends.

Loading plots corresponding to the scores show how the process variables combine to form the scores. **Figure 25** shows the three loading plots of the model of **Figure 24**. The first loading shows that temperature, level, and pH increase during the batch evolution, while ethanol decreases. The second shows that molasses, ammonia, and air have a quadratic behavior increasing and then decreasing during the batch evolution.

2.10.5.2.2 Contribution plots for the interpretation of specific events

The interpretation of interesting points in score plots, Hotelling's T^2 plots, and DModX plots such as outliers, jumps, trend breaks is facilitated by the use of contribution plots. As shown in Section 2.10.3.2.3

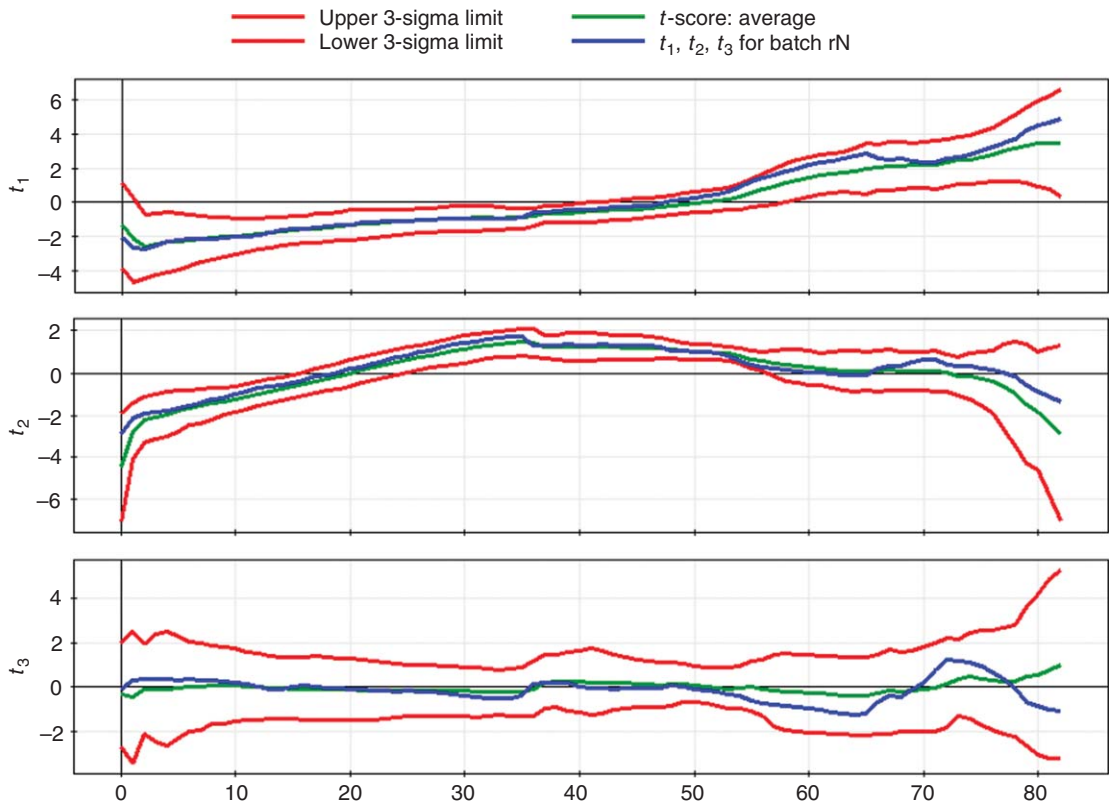

Figure 24 The average score traces (green) for the three components of the yeast observation level model together with the aligned traces of one good batch (N in blue).

these are basically the weighted difference between the **X**-values in the actual point and those of a reference point. Group contributions can also be calculated as the weighted difference between the averages of two groups.

2.10.5.2.3 Nylon example, batch 49

For comparison with the BWU approach (**Figure 22**), the OWU-based Hotelling's $T^{'2}$ diagnostics of the nylon example batch 49 are shown in **Figure 26**.

In the nylon lower OWU model, batch 49 deviates for all three components but not in the DModX (SPE) plot. Hence, the Hotelling's $T^{'2}$ plot provides the best detection of the process problem in batch 49 (**Figure 26**, upper left plot). The contribution plot (right) shows the pressure variables 8 and 9 to deviate the most.

2.10.5.3 Summary of Monitoring Approaches

Monitoring with any of the BWU approaches has several advantages as well as disadvantages. For offline monitoring, only at the end of each batch can the approach take advantage of the ability of the BWU method to capture the dynamic, time-varying covariance structure of batches and to compare differences among batches directly. It can also effectively incorporate initial condition and raw material data (**Z**) and final product quality (**Y**) through multiblock PLS models. Alignment is also not a problem as the complete batch is available to align in the same way as all historical batches were aligned. If the number of variables is large (usually because of the presence of online multivariate sensors), an OWU analysis to first reduce the number of sensor variables may be warranted.

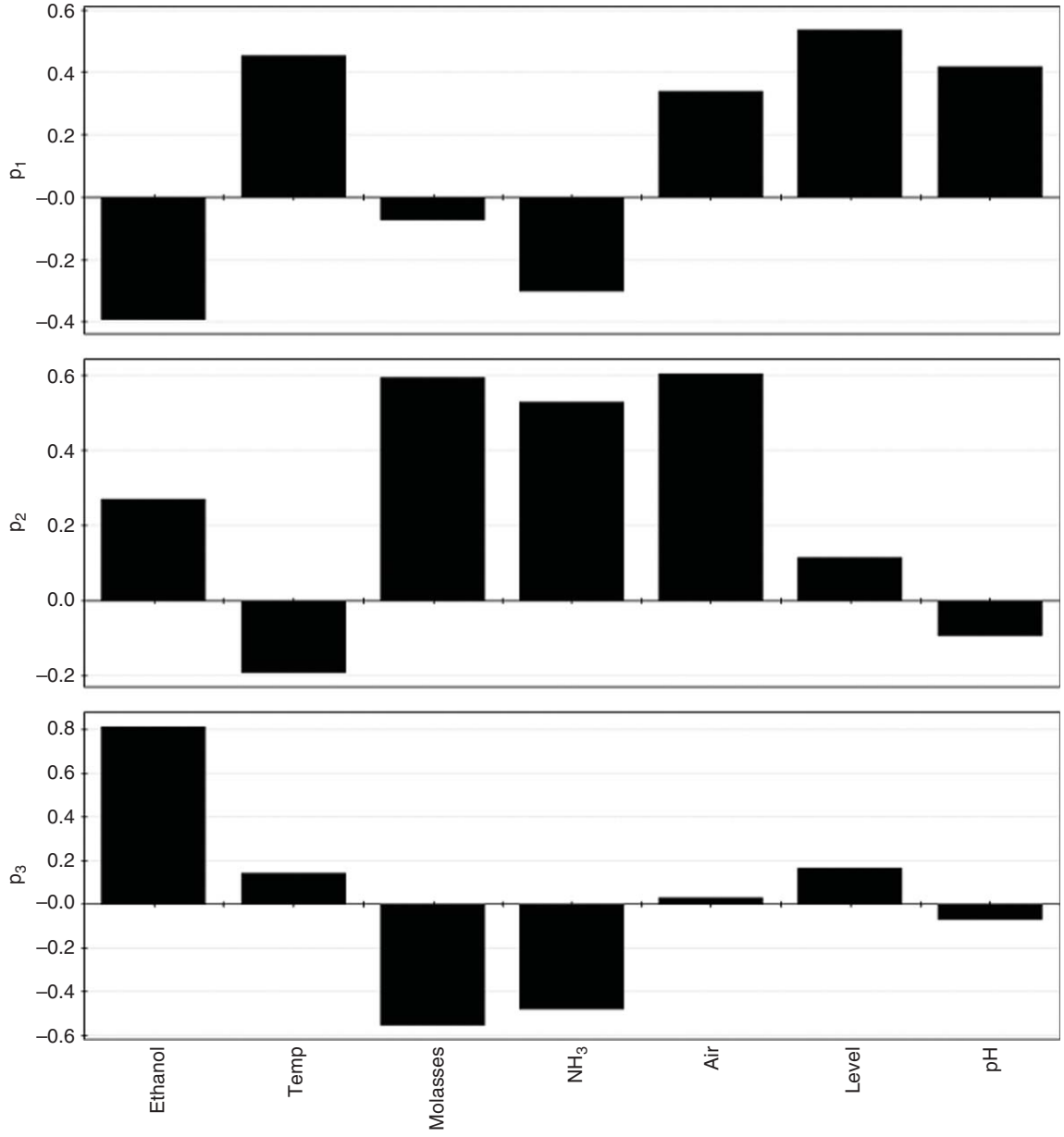

Figure 25 The three loading profiles of the yeast observation level model, p_1, p_2, and p_3 (shown from top to bottom).

If it is desired to detect deviations from normal batch behavior during the progress of the batch, then the online BWU approach or the OWU approach may be used. As illustrated in **Figures 23** and **24**, these approaches provide adequate online monitoring charts. Disadvantages include the need for an indicator or maturity variable in cases where the batch durations vary significantly, and the need for the imputation of the missing data beyond the current time point if BWU is used. OWU monitoring needs no imputation, but at the expense of often requiring multiple models for the different phases to capture the nonlinear time-varying structure of the batches. The BWU approaches, including OWU–TBWU, also allow for monitoring to be based on PLS models for the final quality and direct incorporation of initial conditions ($\mathbf{Z}$) such as raw material properties.

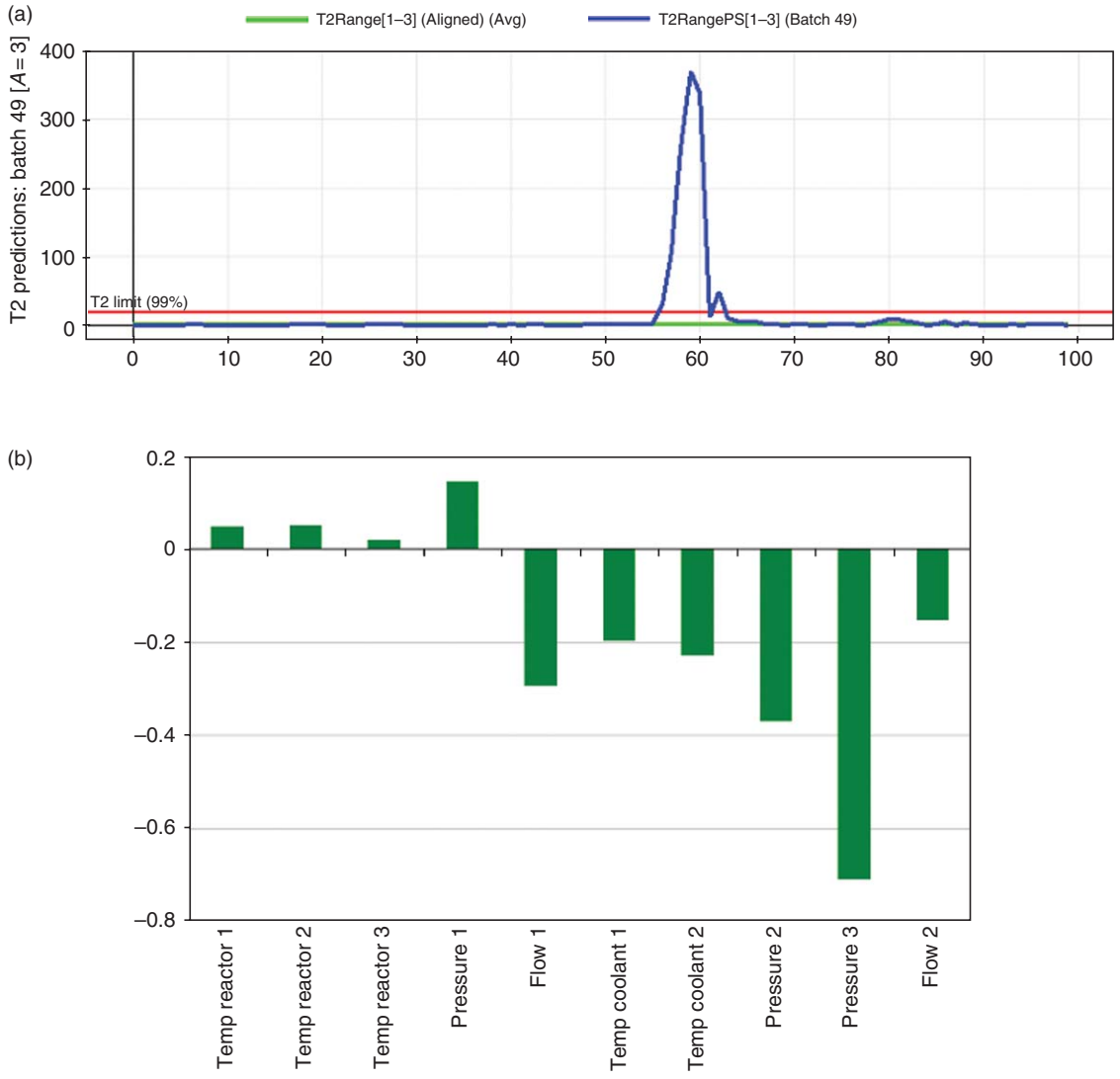

Figure 26 Batch control charts for the OWU model ($A = 3$) of the nylon data and batch 49 (corresponding to **Figure 22**). (a) Hotelling's T^2 (three components). (b) Contribution plot based on the Hotelling's T^2 at point 59. The trace of variable 8 (Pressure 2) is given in **Figure 22** and the trace of variable 9 (Pressure 3) is very similar.

2.10.6 Control and Optimization of Batch Processes

In this section, we provide a very brief overview of the optimization and control of batch processes based on latent variable methods. By *optimization* we mean solving for batch operating variable set-point trajectories and initial conditions and recipes that can achieve a desired set of final product properties, subject to certain constraints on the process. By *control* we mean the adjustment of manipulated variables at certain times during the batch to achieve a desired objective.

2.10.6.1 Control of Batch Processes

There are two levels of batch control problems: (1) the low-level control problem of tracking a specified set-point trajectory on some variables such as temperature, pressure, and pH; and (2) the high-level problem of controlling the vector of final product quality properties to a desired target.

The low-level problem of trajectory tracking is often satisfactorily handled by using simple proportional-integral-derivative (PID) controllers. On the contrary, when these controllers cannot achieve sufficiently tight control because of the type of changes demanded in the set-point trajectories or to the time-varying dynamics throughout the batch (e.g., changing heat transfer and reaction rates), then more powerful model predictive controllers (MPCs) based on either fundamental models or on models based on PCA have been proposed. PCA approaches use either BWU or a combination of OWU and BWU to capture the time-varying dynamics over the whole batch or over windows during the batch and adjust the manipulated variable trajectories at every time point to minimize the deviation of the projected process variables (see Section 2.10.4) from their desired trajectories over some future time horizon.[12]

However, the more important control problem is the upper level control of final product quality and so we will deal exclusively with this problem for the remainder of the control section. To control the final product properties, a BWU-PLS model incorporating the effects of the initial conditions ($\mathbf{Z}$) and process variable trajectories ($\mathbf{X}$) on the final properties ($\mathbf{Y}$) is required. With tightly automated batch processes, many of the batches achieve acceptable final product quality without any intervention or control. Therefore, strategies based on making just one or a few midcourse corrections on only those batches that appear to be out of control are generally sufficient and well accepted by industry. This is analogous to the occasional midcourse corrections used by NASA on space flights to hit a target orbiting around the moon.

As shown in the earlier parts of this chapter, the essential time-varying dynamic covariance structure of the process variable trajectories for each batch and the influence of these trajectories on the final product quality ($\mathbf{Y}$) are well captured by a few latent variables in a BWU-PLS model. Therefore, one can perform control calculations entirely within this reduced dimension latent variable space, a great simplification of the problem.

The midcourse correction strategy consists of selecting one or more decision points (θ_i) in time or in a maturity variable at which control decisions are to be made. Based on all the data up to and including time(s) θ_i, the BWU-PLS model is used to predict the remaining process variable trajectories and final product quality assuming no changes (from nominal behavior) in the manipulated variable trajectories over the remainder of the batch (i.e., under a no-control scenario). These imputations are performed using one of the missing data imputation methods discussed in Section 2.10.4 (see **Figure 21** for an example of this projection). If the projected final product quality is outside of an acceptable multivariate target region,[20] then a correction is called for.

The midcourse corrections may take the form of either a simple instantaneous adjustment such as the amount of a chemical or nutrient addition at time θ_i or an adjustment of the trajectories of the manipulated variables for the remainder of the batch.[20–23] Note that, in addition to using historical operating data to build the model, a few designed experiments are needed to capture the causal effect of the manipulated variable adjustments at times θ_i. A typical objective function (minimized at decision time θ_i using a quadratic programming algorithm) is given by Flores-Cerrillo and MacGregor:[22]

$$\underbrace{\min}_{\Delta t(\theta_i)} (\hat{\mathbf{y}}-\mathbf{y}_{sp})^{\mathrm{T}} \mathcal{Q}_1 (\hat{\mathbf{y}}-\mathbf{y}_{sp}) + \Delta\mathbf{t}^{\mathrm{T}} \mathcal{Q}_2 \Delta\mathbf{t} + \lambda T^2$$

$$\text{such that} \quad \hat{\mathbf{y}}^{\mathrm{T}} = (\mathbf{t} + \hat{\mathbf{t}}_{\text{present}})^{\mathrm{T}} \mathbf{Q}^{\mathrm{T}}$$

$$T^2 = \sum_{a=1}^{A} \frac{(\Delta t + \hat{t}_{\text{present}})_a^2}{s_a^2}$$

$$\Delta\mathbf{t}_{\min} \le \Delta\mathbf{t} \le \Delta\mathbf{t}_{\max} \tag{6}$$

The optimization in Equation (6) computes, at decision time θ_i, the adjustments in the final values of the latent variables, $\Delta\mathbf{t}(\theta_i)$, needed to bring the projected product quality at the end of the batch to its set point, subject to constraints on the magnitude of the score adjustments (note that not all terms in the objective function are necessary). Given these computed optimal score adjustments, the manipulated variables (single variables or trajectories) for the remainder of the batch can then be calculated as given in Equations (2)–(4) in Section 2.10.4.

These calculated manipulated variable trajectories or values will be entirely consistent with the historical operation of the batch as they come from the PLS model built from that data. Therefore, the controller will not call for any unexpected or unusual behavior in the manipulated variable trajectories. Complete details on the

midcourse latent variable control methods with simulation examples are available in other studies.[20–24] A successful industrial application is presented in Yabuki *et al.*[25]

2.10.6.2 Optimization of Batch Processes

The most common problems for the use of data-based latent variable methods for optimization of batch processes are (1) the design of process conditions, recipes, and raw materials necessary to achieve a desired set of final product quality properties[26–28] and (2) the scale-up or transfer of batch operations from one reactor system to another.[26,29,30] Methods for both of these problems have been developed and applied in industry. These approaches are all based on building BWU-PLS models using ($\mathbf{Z}$, $\mathbf{X}$, $\mathbf{Y}$) data from various related products produced in past pilot plant and full-scale batch operations, then specifying a feasible desired product quality (y_{des}), and then inverting these models via an optimization to obtain process trajectory (x) and recipe conditions (z) that are predicted to achieve the y_{des} and at the same time are consistent with historical operations.

2.10.7 Conclusions

This chapter presents a comprehensive overview of chemometric approaches to the modeling, monitoring, and control of batch processes. Various approaches to the modeling of 3D $\mathbf{X}$ arrays that arise in batch data are presented. These include extracting landmark features and applying various unfolding methods, followed by PCA or PLS. These models can then be used to monitor new batches, to provide online predictions of the remaining trajectories and the final product quality, at any time during the batch, and for the control of batch end point and final product quality. All of these are illustrated using data from three industrial data sets. Other batch data examples, not discussed in this chapter, are available in other studies.[3,11,27,31–34]

References

1. Garcia-Muñoz, S.; Kourti, T.; MacGregor, J. F.; Mateos, A. G.; Murphy, G. Trouble-Shooting of an Industrial Batch Process Using Multivariate Methods. *Ind. Eng. Chem. Res.* **2003**, *42*, 3592–3601.
2. Kourti, T.; Nomikos, P.; MacGregor, J. F. Analysis, Monitoring and Fault Diagnosis of Batch Processes Using Multiblock and Multiway PLS. *J. Process Control* **1995**, *5* (4), 277–284.
3. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C.; Wold, S. *Multi- and Megavariate Data Analysis; Principles and Applications*, 2nd ed.; Umetrics AB: Umeå, Sweden, 2006; Vols. 1 and 2.
4. Wold, S.; Kettaneh, N.; Tjessem, K. Hierarchical Multi-Block PLS and PC Models, for Easier Interpretation, and as an Alternative to Variable Selection. *J. Chemom.* **1996**, *10*, 463–482.
5. Wold, S. Cross Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397–406.
6. Nomikos, P.; MacGregor, J. F. Monitoring of Batch Processes Using Multi-Way Principal Component Analysis. *AIChE J.* **1994**, *40* (8), 1361–1375.
7. Nomikos, P.; MacGregor, J. F. Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics* **1995**, *37* (1), 41–59.
8. Nomikos, P.; MacGregor, J. F. Multi-Way Partial Least Squares in Monitoring Batch Processes. *Chemom. Intell. Lab. Syst.* **1995**, *30* (1), 97–108.
9. Wold, S.; Kettaneh, N.; Fridén, H.; Holmberg, A. Modelling and Diagnostics of Batch Processes and Analogous Kinetic Experiments. *Chemom. Intell. Lab. Syst.* **1998**, *44* (1–2), 331–340.
10. Kassidas, A.; MacGregor, J. F.; Taylor, P. A. Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* **1998**, *44* (4), 864–875.
11. Kourti, T. Multivariate Dynamic Data Modelling for Analysis and Statistical Process Control of Batch Processes: Start-Ups and Grade Transitions. *J. Chemom.* **2003**, *17*, 93–109.
12. Flores-Cerrillo, J.; MacGregor, J. F. Latent Variable MPC for Trajectory Tracking in Batch Processes. *J. Process Control* **2005**, *15*, 651–663.
13. Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Comparing Alternative Approaches for Multivariate Statistical Analysis of Batch Process Data. *J. Chemom.* **1999**, *13* (3–4), 397–413.
14. Nelson, P. R. C.; Taylor, P. A.; MacGregor, J. F. Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations. *Chemom. Intell. Lab. Syst.* **1996**, *35* (1), 45–65.
15. Arteaga, F.; Ferrer, A. Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples. *J. Chemom.* **2002**, *16*, 408–418.

16. Garcia-Muñoz, S.; Kourti, T.; MacGregor, J. F. Model Predictive Monitoring for Batch Processes. *Ind. Eng. Chem. Res.* **2004**, *43*, 5929–5941.
17. Yoon, S.; MacGregor, J. F.; Bruwer, M-J. Batch Process Control – When to Add Nutrient Feed and When to Harvest. Presented at the BioProcess International Conference, San Diego, May 2007; available from http://www.prosensus.ca/white-papers/.
18. Adamson, T. Use of PAT Technologies in Semiconductor Manufacturing. ISPE Washington PAT Symposium, IBM, 2006.
19. Nomikos, P. Statistical Control of Batch Processes. Ph.D. Thesis, Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada, 1994.
20. Yabuki, Y.; MacGregor, J. F. Product Quality Control in Semibatch Reactors Using Mid-Course Correction Policies. *Ind. Eng. Chem. Res.* **1997**, *36* (4), 1268–1275.
21. Flores-Cerrillo, J.; MacGregor, J. F. Control of Particle Size Distributions in Emulsion Polymerization Using Mid-Course Correction Policies. *Ind. Eng. Chem. Res.* **2001**, *41*, 1805–1814.
22. Flores-Cerrillo, J.; MacGregor, J. F. Control of Batch Product Quality by Trajectory Manipulation Using Latent Variable Models. *J. Process Control* **2004**, *14*, 539–553.
23. Flores-Cerrillo, J.; MacGregor, J. F. Within-Batch and Batch-to-Batch Inferential Adaptive Control of Semi-Batch Reactors. *Ind. Eng. Chem. Res.* **2003**, *42*, 3334–3345.
24. Flores-Cerrillo, J.; MacGregor, J. F. Multivariate Analysis and Monitoring of Batch Processes Using Batch-to-Batch Information. *AIChE J.* **2004**, *50*, 1219–1228.
25. Yabuki, Y.; Nagasawa, T.; MacGregor, J. F. An Industrial Experience with Product Quality Control in Semi-Batch Processes. *Comput. Chem. Eng.* **2000**, *24* (2–7), 585–590.
26. Jaeckle, C. M.; MacGregor, J. F. Product Design through Multivariate Statistical Analysis of Process Data. *AIChE J.* **1998**, *44*, 1105–1118.
27. Garcia-Muñoz, S.; MacGregor, J. F.; Kourti, T. Optimization of Batch Operating Policies: Handling Multiple Solutions. *Ind. Eng. Chem. Res.* **2006**, *45*, 7856–7866.
28. Muteki, K.; MacGregor, J. F.; Ueda, T. On the Rapid Development of New Polymer Blends: The Optimal Selection of Materials and Blend Ratios. *Ind. Eng. Chem. Res.* **2006**, *45*, 4653–4660.
29. Jaeckle, C. M.; MacGregor, J. F. Industrial Applications of Product Design through the Inversion of Latent Variable Models. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 199–210.
30. Garcia-Muñoz, S.; Kourti, T.; MacGregor, J. F. Product Transfer between Sites Using Joint-Y PLS. *Chemom. Intell. Lab. Syst.* **2005**, *79*, 101–114.
31. Wold, S.; Cheney, J.; Kettaneh, N.; McCready, C. The Chemometric Analysis of Point and Dynamic Data in Pharmaceutical and Biotech Production (PAT) – Some Objectives and Approaches. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 159–163.
32. Antti, H.; Bollard, M. E.; Ebbels, T.; Keun, H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. Batch Statistical Processing of 1-H-NMR Derived Urinary Spectral Data. *J. Chemom.* **2002**, *16*, 461–468.
33. Wiklund, S.; Karlsson, M.; Antti, H.; Johnels, D.; Sjöström, M.; Wingsle, G.; Edlund, U. A New Metabonomic Strategy for Analysing the Growth Process of the Poplar Tree. *Plant Biotechnol. J.* **2005**, *3*, 353–362.
34. Jonsson, P.; Stenlund, H.; Moritz, T.; Trygg, J.; Sjostrom, M.; Verheij, E. R.; Lindberg, J.; Schuppe-Koistinen, I.; Antti, H. A Strategy for Modelling Dynamic Responses in Metabolic Samples Characterized by GC/MS. *Metabolomics* **2006**, *2* (3), 135–143.

Biographical Sketches

Svante Wold is professor emeritus of chemometrics at Umeå University, Sweden, where in 1971 he received his Ph.D. in physical organic chemistry. He is a member of the Swedish Academy of Engineering Sciences, has received many scientific medals and other honors, and has authored and coauthored approximately 400 scientific papers and two books. Approximately 25 graduate students have received their Ph.D. from the Umeå chemometrics group from 1970 through 2005.

In 1987, he was one of the founders of the company Umetrics in Umeå, and after 1990 he has spent much of his time in this company where he now is VP R&D, helping to implement chemometrics methodology in the Simca and Modde software and teach and consult with chemical, pharmaceutical, and other industry all over the world.

Nouna Kettaneh-Wold received her M.Sc. in Physics and Applied Mathematics at the University of Lyon, France. She has several publications in the chemometrics literature on the design of experiments, mixture modeling, and batch process modeling. She is one of the founders, and now general manager of Umetrics, having offices in Sweden, the United States, and the United Kingdom. Umetrics is part of MKS Instruments since 2006 and develops Multivariate Analysis software – the Simca suite – for offline modeling as well as online monitoring and fault detection, and the Modde software for experimental design. Umetrics also provides support and consulting services to industries in these areas.

John MacGregor received his Ph.D. degree in Statistics, MASc degrees in both statistics and chemical engineering from the University of Wisconsin, and a BEng degree in chemical engineering from McMaster University. After several years with Monsanto Company in Texas, he joined the Department of Chemical Engineering at McMaster University in 1972. He currently holds the title of 'Distinguished University Professor' and the Dofasco Chair in Process Automation and Information Technology. He has made significant research and industrial contributions in the areas of advanced control, chemometrics, and polymer reaction engineering.

He is a Fellow of the American Statistical Association, and has received many awards for his work in applied statistics and chemometrics, among them, the Shewhart Medal and the W.G. Hunter Award from the American Society for Quality, and the Herman Wold Medal from the Swedish Chemical Society. He has also received numerous awards from engineering societies, among them, the Century of Achievement Award and the Industrial Practice Award from the Canadian Society for Chemical Engineering, the Computing and Systems Technology Award from the American Institute of Chemical Engineers, and the Guido Stella Award from the World Batch Forum. He is a Fellow of the Royal Society of Canada and the Canadian Academy of Engineering.

Kevin Dunn has a background in Chemical Engineering from the University of Cape Town (undergraduate) and McMaster University (Masters). He helped start ProSensus with John MacGregor and Honglu Yu in 2002, while working as a research engineer at McMaster University. He is currently a full-time employee at ProSensus and is working on projects related to data and image analysis across various industries, including the foods, pharmaceutical, and materials industry.