JOURNAL OF CHEMOMETRICS
J. Chemometrics **13,** 397–413 (1999)

COMPARING ALTERNATIVE APPROACHES FOR MULTIVARIATE STATISTICAL ANALYSIS OF BATCH PROCESS DATA

JOHAN A. WESTERHUIS,† THEODORA KOURTI* AND JOHN F. MACGREGOR
McMaster Advanced Control Consortium, Department of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada

SUMMARY

Batch process data can be arranged in a three-way matrix (batch × variable × time). This paper provides a critical discussion of various aspects of the treatment of these multiway data. First, several methods that have been proposed for decomposing three-way data matrices are discussed in the context of batch process data analysis and monitoring. These methods are multiway principal component analysis (MPCA)—also called Tucker1—parallel factor analysis (PARAFAC) and Tucker3. Secondly, different ways of unfolding, mean centering and scaling the three-way matrix are compared and discussed with respect to their effects on the analysis of batch data. Finally, the role of the time variable in batch process data is considered and methods suggested to predict the per cent completion of batch runs with unequal duration are discussed. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: batch process data; multiway PCA; PARAFAC; Tucker3; Tucker1; batch data alignment

1. INTRODUCTION

Batch and semi-batch processes are characterized by a prescribed processing of materials for a finite duration. Specialty polymers, pharmaceuticals and biochemical materials are produced in batch reactors. Batch annealing in the steel industry, injection molding of polymers, batch distillation and batch etching of silicon wafers are other examples of batch processes. Typical data from batch processes include time-varying trajectories of all the measured process variables collected for a number of batches. Process variables such as temperatures, reactant feed rates, pressure, agitation intensity, etc., measured at many times throughout the duration of each batch, contain valuable information that subsequently can be used to analyze the process behavior, improve its performance and establish acceptable operation limits for process monitoring. For a few processes, product quality measurements are available during production (e.g. density, latex particle size). However, in most cases the quality variables are only available at the end of the batch and cannot be used for on-line

* Correspondence to: T. Kourti, McMaster Advanced Control Consortium, Department of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada. E-mail: kourtit@mcmaster.ca.
† Current address: Chemical Engineering Department, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, The Netherlands.

CCC 0886–9383/99/040397–17 $17.50

Received 30 July 1998
Accepted 19 March 1999

monitoring purposes.

The use of multivariate statistical projection methods for the analysis, on-line monitoring and fault detection and diagnosis of batch processes has been described by Nomikos and MacGregor.[1–3] Tutorials and extensions of these methods for fault diagnosis and for the incorporation of additional information such as properties of the raw materials are given by Kourti and MacGregor[4,5] and Kourti *et al.*[6]

Batch process data can be arranged in a three-way matrix. Nomikos and MacGregor[1–3] showed how to use multiway principal component analysis (MPCA)[1–3] and multiway partial least squares (MPLS)[7] for the analysis of three-way batch data. In MPCA the three-way data matrix is unfolded to a two-way matrix which is analyzed using standard PCA. Other methods that analyze three-way data, such as parallel factor analysis (PARAFAC)[8–11] and Tucker models,[10,12] have been applied in chemistry and are now being suggested for the treatment of batch data.[13] This paper examines some issues related to the use of these three methods in batch process data analysis, but it is not meant to be a comprehensive treatment of the topic. When using unfolded PCA, Nomikos and MacGregor[3] concluded that a specific way of unfolding in which the direction of the batches is maintained is most suitable for batch process monitoring. They also suggested appropriate ways for centering and scaling the data. However, recently some other ways of unfolding, centering and scaling have been used.[14,15] The impact that these kinds of preprocessing will have on the analysis of batch data is examined. Finally, in many situations, batch runs are of different durations. Methods for aligning the batch data and suggestions for predicting the per cent completion of a batch are discussed.

2. COMPARISON OF THREE-WAY METHODS FOR BATCH DATA ANALYSIS

In this section, methods developed to model three-way data are discussed with respect to their use for batch process data analysis. These methods are multiway principal component analysis (MPCA)—or Tucker1—PARAFAC and Tucker3.

2.1. Theory on methods

Typical data from a batch process consist of J process variables measured at K points in time for I batches. In MPCA as used by Nomikos and MacGregor,[1–3,16] the three-way matrix $\underline{\mathbf{X}}(I \times J \times K)$ is unfolded to a two-way array so that the direction of the batches is maintained. Then PCA is performed on the two-way data set. With MPCA, in each component the data are decomposed into a vector in the batch mode, $\mathbf{a}^{\mathrm{M}}$ (the term scores is usually used for this mode in batch analysis), and a matrix $\mathbf{P}_d$ that describes the variation in the combined variable/time space for each component. The loading vectors $\mathbf{P}_d$ ($JK \times 1$), $d = 1, 2, \ldots, D$, can then be refolded into a three-way loading array $\underline{\mathrm{P}}$ ($D \times J \times K$) as shown in Figure 1. The structural model is

$$\hat{x}_{ijk} = \sum_{d=1}^{D} a_{id}^{\mathrm{M}} P_d(j,k) \tag{1}$$

The index M on $\mathbf{a}^{\mathrm{M}}$ denotes a component from MPCA. The components are orthogonal and can be calculated sequentially. Wold *et al.*[7] considered further decomposition of the loading matrix $\mathbf{P}_d$, without centering and scaling, into principal components $\mathbf{c}^{\mathrm{M}}$ and $\mathbf{b}^{\mathrm{M}}$. This feature is discussed later in Section 2·2 (Table 1).

Other decomposition methods such as PARAFAC[8–11] and Tucker3[10,12] have also been used in chemometrics for the analysis of three-way data. A tutorial on PARAFAC is given by Bro.[11] Figure 2 shows the decomposition of the three-way matrix into trilinear component loadings ($\mathbf{a}_f^{\mathrm{P}}$, $\mathbf{b}_f^{\mathrm{P}}$ and $\mathbf{c}_f^{\mathrm{P}}$)

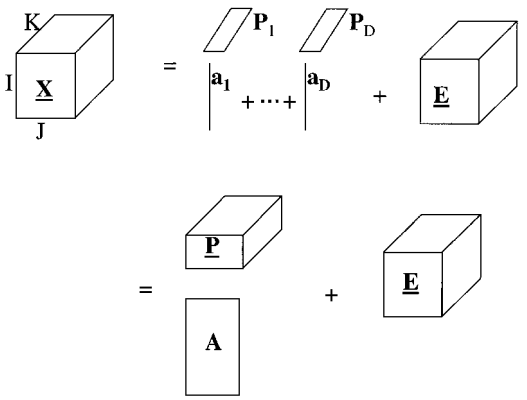

Figure 1. Decomposition of X by MPCA to D components. The X matrix contains data on J variables measured at K points in time for I batches. E is the matrix of the residuals

according to a PARAFAC model. The three-way matrix X can be expressed as the sum of several such trilinear components. E is the matrix of the residuals. The structural model of PARAFAC is given by

$$\hat{x}_{ijk} = \sum_{f=1}^{F} a_{if}^{\mathrm{P}} b_{jf}^{\mathrm{P}} c_{kf}^{\mathrm{P}} \tag{2}$$

where F is the number of components. **P** is the index on the vectors $\mathbf{a}^{\mathrm{P}}$, $\mathbf{b}^{\mathrm{P}}$ and $\mathbf{c}^{\mathrm{P}}$ used for PARAFAC.

PARAFAC models have the same dimension in the three modes, but this is not necessarily the case in Tucker3 models. Figure 3 shows the decomposition according to a Tucker3 model. A Tucker3 model decomposes the three-way data into loadings in each direction ($\mathbf{A}_{\mathrm{T}}$, $\mathbf{B}_{\mathrm{T}}$ and $\mathbf{C}_{\mathrm{T}}$) which are connected through a core matrix G. Here G is a $D \times E \times F$ matrix, **A** is $I \times D$, **B** is $J \times E$ and **C** is $K \times F$. When one of the dimensions in G is the same as in X (e.g. $D = I$, $F = K$ or $E = J$), then we have a Tucker2 model. If two dimensions are the same (e.g. $F = K$ and $E = J$), then we have a Tucker1 model and this is equivalent to unfolding X into a two-way matrix and performing PCA on it. The

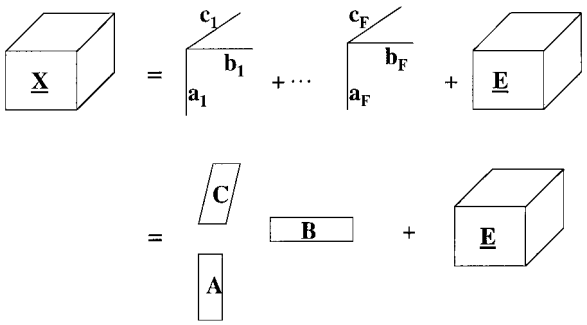

Figure 2. Decomposition of X by PARAFAC to F components

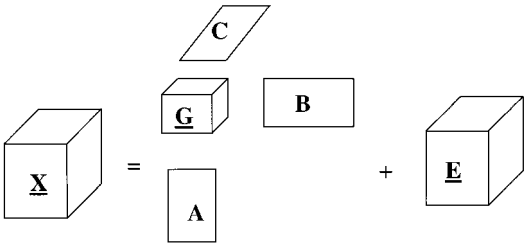

Figure 3. Decomposition of $\underline{X}$ by Tucker3 model

structural model of Tucker3 is

$$\hat{x}_{ijk} = \sum_{d=1}^{D} \sum_{e=1}^{E} \sum_{f=1}^{F} a_{id}^{T3} b_{je}^{T3} c_{kf}^{T3} g_{def} \tag{3}$$

Here we use T3 as the index on the vectors $\mathbf{a}^{T3}$, $\mathbf{b}^{T3}$ and $\mathbf{c}^{T3}$ for the Tucker3 model. Notice that the PARAFAC model is a Tucker3 model with $D = E = F$ where the core has all the elements zero except for the superdiagonal $g_{111}, g_{222}, \ldots, g_{FFF}$.

An introduction to the Tucker3 models together with some simplifications is given by Geladi.[12] For the PARAFAC and Tucker3 models the components are calculated simultaneously. Therefore, unlike MPCA, in PARAFAC and Tucker3 the first batch loading of a two-component model may be different from the first loading of a one-component model. Kiers[17] shows that PARAFAC can be considered a constrained version of Tucker3, and Tucker3 a constrained version of Tucker1 (MPCA).

In the following subsection we will compare the potential of the various three-way methods for the analysis of batch data from two points of view. First, the main results will come from comparing the abilities of the methods to model two previously published sets of data from batch polymerization processes.[1,3] Secondly, we will also attempt to provide some insight into the nature of batch data and the suitability of the methods to capture this structure.

2.2. Case studies

The methods will be compared using data from two batch processes.

- The first data set from an industrial batch polymerization process was supplied by DuPont.[3] Data from 36 good batches were used to build the models. Each batch had a duration of 100 time intervals, and ten process variables were monitored throughout the process.
- The second data set is from a simulated polymerization of styrene–butadiene (SBR).[1] Here 51 good batches were used and six variables were monitored at 200 time intervals. Three other noisy process variables in this data set were not used.

For this first analysis the same data preprocessing was used for each method. The three-way data were unfolded in such a way as to maintain the batch direction (unfolding **D** or **E** in Figure 5), and each column was centered to zero mean (i.e. the mean trajectories of the process variables were removed) and scaled to unit variance. Other forms of unfolding, centering and scaling and their consequences will be discussed later in Section 3. The calculations for the PARAFAC and Tucker3 models were performed using the *N*-way toolbox of Andersson and Bro.[18] The MPCA calculations were performed using the BatchSPC software of the McMaster Advanced Control Consortium. For

Table 1. Cumulative per cent variation explained with PARAFAC, Tucker3 and MPCA models. The dimensions of the Tucker3 models (batch, variable, time) are given in parentheses. The cumulative per cent variation explained by PCA of the loading matrix $\mathbf{P}_d$ is also given

	Number of components	PARAFAC	Tucker	MPCA	PCA of $\mathbf{P}_d$[a]
DuPont data	1	21	27 (1,2,2)	39	53, 69, 82
	2	30	32 (2,2,2)	50	74, 85
	3	34	40 (3,3,3)	57	46
SBR data	1	19	20 (1,2,2)	21	90, 96, 100
	2	28	28 (2,2,2)	33	63, 92, 99
	3	32	34 (3,3,3)	40	55, 84, 97

[a] Number of components determined by cross-validation.

the PARAFAC loadings in the batch direction, orthogonal constraints are applied. These loadings will be used to build monitoring schemes, and orthogonality allows for monitoring each score separately. Furthermore, PARAFAC models without orthogonal constraints were sometimes found to produce degenerate solutions. Degenerate solutions are often characterized by highly correlated loadings, and the resulting models are often unstable and unreliable.[11]

Table 1 shows the results of modeling the DuPont and SBR batch process data with PARAFAC, Tucker3 and MPCA. Columns 3–5 show the cumulative percent variation explained for one, two and three components. Tucker (1,1,1) would have been a PARAFAC, so we chose a Tucker (1,2,2) to compare models with only one component in the batch direction but show the effect of more components in the variable and time directions; any number of components in the variable and time directions could have been used. Notice that when the models are compared on an 'equal component basis', MPCA explains the highest percentage of variation, followed by Tucker3 and PARAFAC. For the DuPont data, for example, with three components, 57% is explained by MPCA, 40% by a Tucker3 (3,3,3) and 34% by PARAFAC. This illustrates a behavior that is theoretically expected because of the structure of the models. A comparison on more 'optimal' models (cross-validated models) for each case will follow, after we discuss another feature of this table.

The last column in Table 1 shows results from a PCA performed on the loading matrix $\mathbf{P}_d$ obtained from MPCA as illustrated in Figure 4. It gives the cumulative percentage of variation explained by performing a PCA on the $\mathbf{P}_d$ matrices using only cross-validated components. The first line (DuPont data) of this column reads that if a PCA is performed on the first-component loading matrix $\mathbf{P}_1$, then one component explains 53% of variation in $\mathbf{P}_1$ and there are three components explaining a total of 82% of $\mathbf{P}_1$. (Only three components were important by cross-validation.) Notice now that 53% of 39

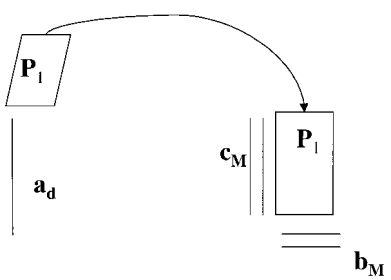

Figure 4. PCA can be performed on $\mathbf{P}_1$ matrix

Table 2. Per cent variation explained by fitting (R^2) and cross-validation (Q^2) for different scalings of $\underline{X}$. The initial sum of squares for each different scaling is shown at the bottom of the table

	Autoscaling		Variable scaling	
	R^2	Q^2	R^2	Q^2
PARAFAC (5)	39	37	40	36
Tucker3 (4,5,3)	44	41	46	43
Unfold PCA (3)	57	48	71	63
Initial sum of squares	3500		35900	

(which is the percentage explained by MPCA) equals 21 (the percentage explained by the PARAFAC model). Notice also that the percentage explained by the Tucker (1,2,2) model (27%) equals the fraction of $\mathbf{P}_1$ explained by two PCs times the percentage of variation explained by the MPCA model ($0{\cdot}69 \times 39 = 27$). The same feature was also found to be true for the SBR data. Notice that this feature is only observed for the first dimension $\mathbf{P}_1$ and it is not known if it is general for all data sets.

This observation would also imply that for a one-component model, only if the MPCA loading matrix $\mathbf{P}_1$ were of rank one ($\mathbf{P}_1 = \mathbf{c}_1^{\mathrm{M}}(\mathbf{b}_1^{\mathrm{M}})^{\mathrm{T}}$) would the PARAFAC model explain the same percentage of variation as MPCA. However, the nature of batch data is such that one would normally expect strong interactions between the time and variable dimensions, which would result in $\mathbf{P}_1$ being of rank higher than one. Batch and semi-batch process data consist of several variables that are measured in time. These process variables are not only correlated to each other at any given time, but also correlated to events that happen at earlier times in the process (hence their three-way nature). Furthermore, the nature of the correlation among the variables and the nature of their correlation with earlier time periods will, in most batch processes, constantly change over the course of the batch. This is because batch processes are usually operated to proceed in several stages, where some variables are deliberately manipulated in a different manner in various stages in order to accomplish different objectives. For example, in the DuPont process the major phenomenon occurring early in the batch is the vaporization of solvent, and the reactor heating medium flows are adjusted to provide a desired rate of change of temperature and control of pressure. However, part way through the batch the operating procedure is changed and the long-chain polymerization reactions begin. The temperatures and pressures during this second stage are correlated with one another in a very different manner than during the first stage, and furthermore, their correlation with prior history is very different. We refer to this as interactions between the time and variable dimensions. These interactions lead to MPCA loading matrices $\mathbf{P}_d$ that usually have ranks much higher than one.

Thus the nature of three-way batch data is often quite different from the trilinear PARAFAC decomposition, where for each dimension (f) the time direction loading parameters (c_{kf}) are the same for all variables at each time point k and the variable dimension loading parameters (b_{jf}) are the same over all times for each variable j. As a result, PARAFAC is not capable of modeling such complex interactions within each of its individual components, and these can only be captured, if at all, by using several components.

The results in Table 1 only considered the per cent variation explained in the fitted data (R^2) by maintaining the same number of components in the batch direction for each model. No attempt was made to select an 'optimal' number of components for each model dimension. In Table 2 we rectify this by using models based on some 'optimal' selection and comparing them on the DuPont data with respect to the percentage explained by fitting (R^2) and the percentage explained by cross-validation (Q^2). Based on a leave-one-batch-out cross-validation procedure, a PARAFAC model with five components and an MPCA model with three components were selected. A Tucker3 model with four

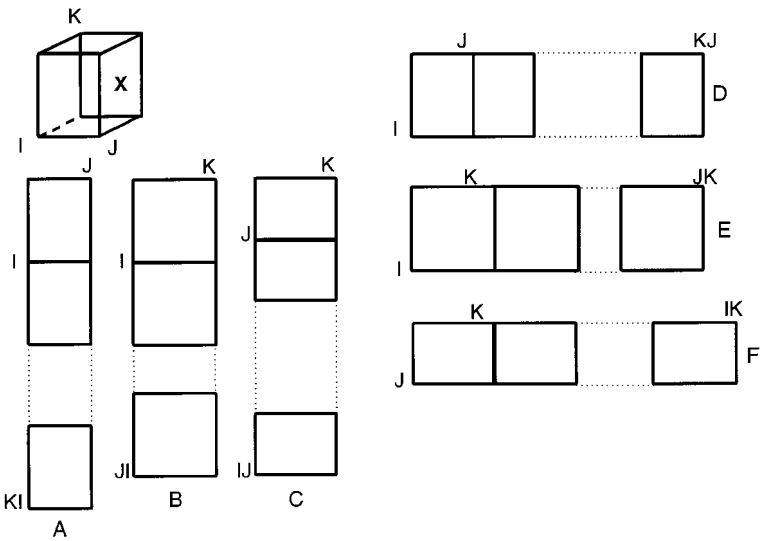

Figure 5. Unfolding three-way data cube $\underline{X}$ into six different two-dimensional matrices

components in the batch direction, five in the time direction and three in the variable direction was selected. These numbers come from a PCA analysis on each of the unfolded matrices ($I \times JK$, $J \times IK$, $K \times IJ$) and selection of the number of components from scree plots in each case. (In the batch direction it was unclear whether three or four components were needed, so four were chosen.)

Column 2 of Table 2 gives the results when the batch data were mean centered in the batch direction and autoscaled. From the table it can be seen that the MPCA model explains the greatest percent variation in the fitted data (R^2), followed by Tucker3 and PARAFAC. This is expected, since MPCA is the most flexible model. This has been pointed out by Bro,[19] who states: 'A Tucker1 model always fits data better than a Tucker3 model which again will fit better than a PARAFAC model, all except for extreme cases where the models may fit equally well'. However, the interesting thing in Table 2 is that the percentage explained by cross-validation predictions (Q^2) is also clearly the best for the MPCA model, followed again by Tucker3 and PARAFAC. Notice that these conclusions are only with respect to the model fits and the predictions via cross-validation. Further studies as to the potential of the models for process analysis and monitoring need to be performed. Table 2 also shows results for two ways of scaling $\underline{X}$. The scaling issue is discussed in the next section.

3. UNFOLDING, MEAN CENTERING AND SCALING

Typical data from a batch process consist of J process variables measured at K points in time for I batches. In order to be able to use the standard PCA and PLS methods, the three-way data matrix has to be unfolded to a two-way matrix. Figure 5 shows that the three-way data matrix $\underline{X}$ ($I \times J \times K$) can be unfolded in six different ways. This results in the following two-dimensional matrices: **A** ($KI \times J$), **B** ($JI \times K$), **C** ($IJ \times K$), **D** ($I \times KJ$), **E** ($I \times JK$) and **F** ($J \times IK$). However, for PCA and PLS, matrices **B** and **C** are equal and matrices **D** and **E** are equal in that they are the same matrix with just the rows or columns rearranged. Matrix **F** is the transpose of **A**, and a PCA would just switch the scores and loadings of the two matrices if no centering or scaling was applied. Each of the different rearrangements of the three-way data matrix $\underline{X}$ into a large two-dimensional matrix followed by PCA on this matrix corresponds to looking at a different type of variability.

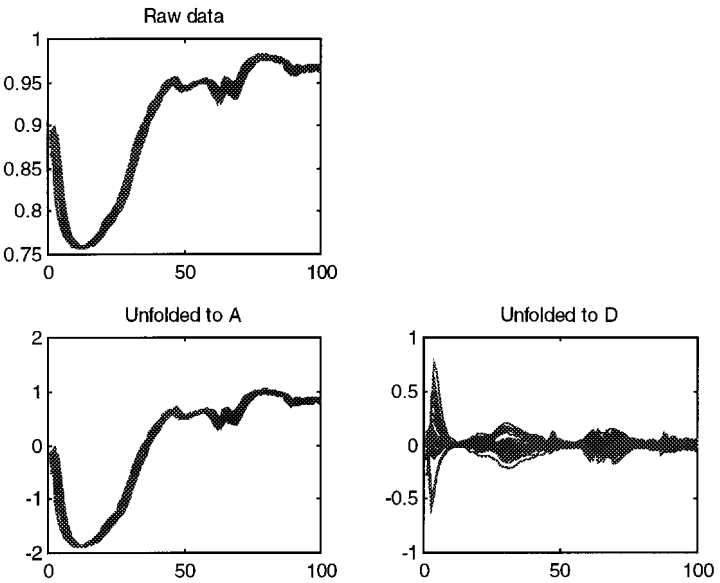

Figure 6. Raw data of variable 2 (trajectories of 36 batches) of DuPont data set unfolded to matrix **A** and to matrix **D** and mean centered afterwards

3.1. MPCA

In the batch process monitoring methods introduced by Nomikos and MacGregor,[1–3], [16] the three-way matrix with process data is unfolded to matrix **D**. This is the most meaningful way of unfolding for batch analysis and monitoring. The whole batch (thus all process variable trajectories obtained during the batch) is considered as one object. With this kind of unfolding, one studies the differences between the batches. It allows comparing each batch separately against a group of good batches to classify it as a good or a bad batch. It also gives the possibility to relate the measurements of each batch to quality measurements of the final product. Mean centering matrix **D** removes the mean trajectories of all process variables, thereby removing the main non-linear and dynamic component from the data. A PCA performed on these mean-corrected data is therefore a study of the systematic variation in all the variable trajectories about their mean trajectories. If a new batch is compared with the PCA model of good batches, the variation should be consistent with the systematic variation of the PCA model.

Recently, applications have been presented in which the three-way matrix with process data was unfolded to matrix **A** and scaled to zero mean and unit variance.[14,15,20] With this way of unfolding, each point of the trajectory of every batch is considered an object. Mean centering matrix **A** simply subtracts a constant (the grand mean of each variable over all batches and all times) from the trajectory of each variable in each batch. This leaves the non-linear time-varying trajectories in the data matrix. This type of unfolding and mean centering may be useful when a wide range of different trajectories have been tried for some of the variables (i.e: when, for the same variable, trajectories obtained for different batches do not fluctuate around the same mean). Such an example is the case of designed experiments where it is desired to find what variation in the trajectory shapes is correlated with various quality variables.[21] However, it is not really suitable for batch monitoring or for the analysis of batches that have been obtained in a consistent manner, as will be illustrated below.

Figure 6 shows the raw data of one of the ten process variables for the 36 batches of the DuPont

Table 3. MPCA results for different ways of unfolding matrix $\underline{X}$. The number of components required to explain the same variation in the process data is much larger when the data are unfolded to **A** and mean centered than to **D** and mean centered

	DuPont data (10 variables)	SBR data (9 variables)
Total sum of squares (SS_{tot})	25336	$1{\cdot}68 \times 10^{10}$
Sum of squares **A** centered ($SS_{\mathbf{A}}$)	2201	$1{\cdot}03 \times 10^{8}$
Sum of squares **D** centered ($SS_{\mathbf{D}}$)	26·3	$2{\cdot}77 \times 10^{6}$
% expl. SS **D** with # PCs	57% with 3 PCs	30% with 3 PCs
% SS needed **A** with # PCs	99·5% with 6 PCs	98·1% with 6 PCs

data unfolded both to matrix **A** and to matrix **D** and mean centered afterwards. The dynamic trajectory is still present in the data after mean centering matrix **A**, whereas after mean centering matrix **D**, only the variation between the batches remains.

A PCA on the unfolded matrix **A** needs many more components to describe the same systematic variation in the data as a PCA on the unfolded matrix **D**. In fact, it usually needs almost as many components as the number of variables in the original matrix **A** to end up with the same residual sum of squares as a cross-validated PCA of matrix **D**. This is logical if the relationships between the variables change over the course of the batch. This will be the case if some of the variables are being manipulated at various times to force the reaction to achieve certain results. To illustrate these points, we again use the DuPont and SBR batch data from the previous section. The only difference is that for the SBR data set all nine variables are used here instead of the six mentioned before, so that we can compare it with the work already presented by Nomikos and MacGregor.[1]

Table 3 shows results from the two data sets. The first row gives the total sum of squares for each data set before mean centering (SS_{tot}). The next two rows give the sum of squares left after mean centering the data unfolded to **A** ($SS_{\mathbf{A}}$) and to **D** ($SS_{\mathbf{D}}$). Then the percentage of sum of squares of matrix **D** explained by PCA is given (% expl. SS **D**), together with the number of components (PCs) needed to explain such variation. Finally we give the percentage of sum of squares of matrix **A** (% SS needed **A**) that it is necessary to explain in order to end up with the same residual sum of squares as obtained with unfolding and centering via **D**, together with the number of components required to explain this variation. For the DuPont data set the total sum of squares of the raw data of all variables is 25 336. After mean centering, the sum of squares is 2201 for matrix **A** and only 26·3 for matrix **D**. For the **D** matrix, three principal components were selected that explained 57% of the variation in the process data,[3] so the residual sum of squares equals 11·35. This means that 99·5% of the variation has to be explained if the three-way matrix is unfolded to matrix **A** to end up with the same residual sum of squares. Six PCs are necessary to describe this percentage. For the SBR data, three PCs explained only 30% of the variation in the process variables.[1] To end up with the same residual sum of squares for matrix **A**, six PCs are needed. In both the DuPont and SBR data sets the first scores from matrix **A** serve mainly to describe the mean trajectories of the variables in time, and the last few components describe most of the important variation among the batches.

Hence, for purpose of process monitoring, unfolding and mean centering in the manner (matrix **A**) offers very little benefit. It focuses on the wrong source of variation in the data, namely the mean trajectories of the variables over all batches, rather than what is of greatest interest—the variation among the batches (i.e. about the mean trajectories). Furthermore, in many situations (as those illustrated here) it gives very little reduction in the dimension of the problem, needing almost as many latent variables as original variables in order to describe all the structured variation in the trajectories.

Table 4. Cumulative per cent variation described with PARAFAC model for data obtained by unfolding $\underline{X}$ to matrix **A** and then centering and autoscaling columns of **A**

Number of components	DuPont data	SBR data
1	65	56
2	88	83
3	96	96
4	98	
% needed to account for removing average trajectories	99	98

3.2. PARAFAC

Wise[15] used PARAFAC to model the three-way batch process data of a semiconductor etching process for monitoring purposes. As a preprocessing step, the three-way matrix was unfolded to matrix **A**, autoscaled to zero mean and unit variance and folded back into the three-way shape. As a result, the mean of each column in the batch direction of the three-way matrix is not zero. Normally this is not a problem if one is interested in the trajectory of the variables. In batch process monitoring, one is not interested in the trajectories of the process variables but in the deviations from their corresponding mean values. Therefore in that case it is important to remove the mean trajectories of the process variables before the PARAFAC analysis. Just as was shown for the MPCA case, also a large number of PARAFAC components are necessary to describe the mean trajectories of the process variables, if these mean trajectories are not removed.

Table 4 shows the per cent variation explained with a PARAFAC model when the data are unfolded to **A**, autoscaled to zero mean and unit variance and folded back for the two data sets. For the DuPont data the total sum of squares of the data after scaling and centering in this manner (i.e. **A**) is 35 990. A four-component PARAFAC model on these data explains 98% of this variation (Table 4). However, by simply removing the mean trajectory of each variable (mean centering using unfolding **D**), the sum of squares is only 351, a 99% reduction. Therefore the four-component PARAFAC model would be missing much of the batch-to-batch variation that it is desired to monitor. As seen in Table 4, similar results were also observed for the SBR data.

Figure 7 shows the trajectories of two process variables of the 36 batches from the DuPont data when the mean trajectories were not removed from the data. The top figures show the trajectories of the raw data, and the following figures show the trajectories of the residuals (E1–E4) after the whole DuPont data set is modeled with one, two, three and four PARAFAC components. Still after four PARAFAC components the mean at each time point is not equal to zero. This means that the mean trajectory of the variables is still not completely removed, even if 99% and 98% of the sum of squares of the two variables have been described respectively. Therefore the main part of the systematic variation between the batches has not been modeled and will still be in the residuals.

It was also observed that if orthogonal constraints are used in the case of non-zero means in the batch direction, a large part of the variation in the data could not be described.

3.3. Scaling batch data before using the multiway methods

It has been argued by several authors that autoscaling is not the proper scaling for three-way matrices because it destroys any trilinear structure of data and incorporates additional three-way components in some models. It has also been argued that although batch data are not trilinear, using autoscaling

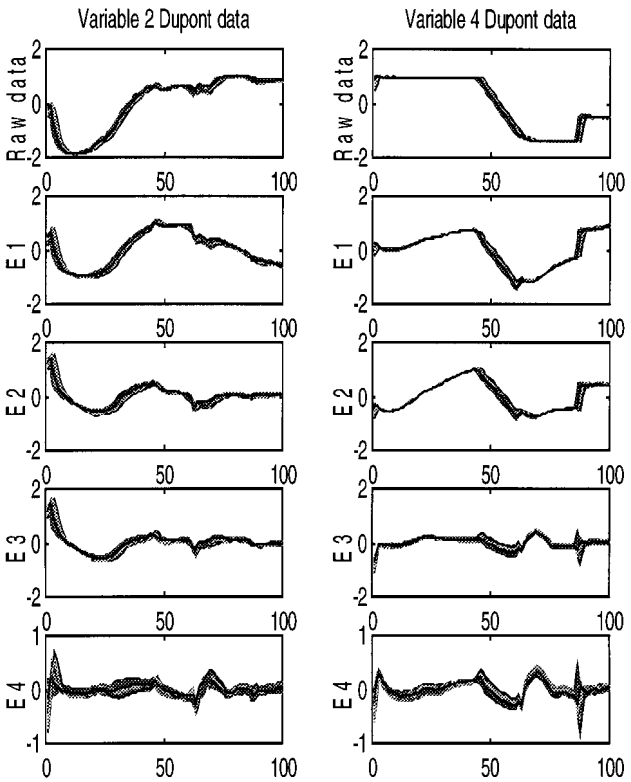

Figure 7. Raw data of two variables from DuPont data, and their corresponding residual trajectories after one, two, three or four PARAFAC components have been modeled

will have a bad influence if a PARAFAC model is used to decompose the three-way matrix. It has also been argued that a proper scaling is variable scaling in which each variable has equal variance. We believe that the scaling of the available data should be chosen in a way which is most appropriate to the problem one is trying to solve, and not such that will bias the results towards one method. Once the problem is correctly formulated, we should then seek the method to solve it.

Nomikos and MacGregor[3] report that they tried both forms of scaling in MPCA with no substantial difference in the results on the DuPont data. The number of components needed for modeling was the same in both cases, as was the fault detection capability. Also, Nomikos[21] states that for the data sets he used: 'another way of scaling which gives similar results to what we use in this work is to scale each variable by its overall (throughout batch duration) standard deviation. The benefit from such scaling is that periods with more variability will be weighted more and will have a greater influence on the MPCA model'. He continues by warning: 'But, if the variability in a particular period is very large, this will result in the rest of this variable's history being *ignored* in the MPCA model'.

Figure 8 shows the raw data and the data after mean centering and different types of scaling. With autoscaling, all the periods during the variable trajectory are treated equally. If the data are scaled by variable, then periods of high variability in each variable's trajectory will be weighted more than the rest. In other words, periods with a lot of noise in a specific variable (e.g. the start-up of the process in Figure 8 c, time 0–20) are weighted more than periods with consistent but highly structured variability. However, in batch process monitoring, it is the deviations in the structured variability that

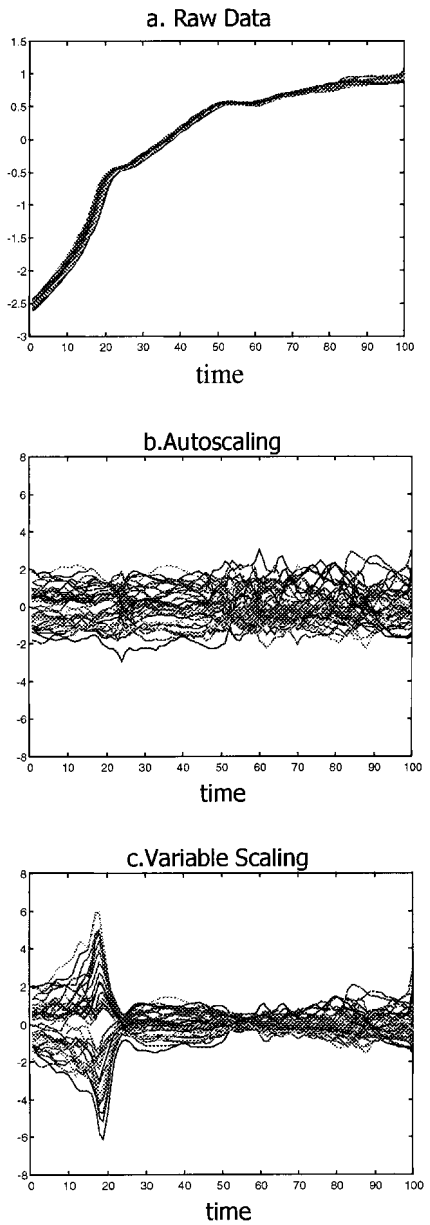

Figure 8. Raw data on trajectories of a variable for 36 batches, and trajectories plotted after mean centering and applying different scaling on same variable

we wish to detect. If the periods that do not vary a lot get a small weight and we emphasize periods with high uncertainty, then it may not be possible to achieve our objective. Therefore scaling should always be related to the objectives of the problem. Based on the experience of two of the authors (T.K. and J.F.M.), autoscaling will be preferable for batch process monitoring if there is no prior knowledge of the process behaviour and the type of faults that may occur. However, if there is prior

knowledge that a specific period during the process is critical, then more weight should be given to this period, or sampling of the data should become more frequent during this period. The latter approach is another way to give higher weight to variables for a certain period.

Table 2 shows the results on 'optimal' models with both types of scaling. For the DuPont data the type of scaling had no effect on the number of components and on the fact that MPCA has higher R^2 and Q^2 than Tucker3 and PARAFAC.

4. THE TIME VARIABLE IN BATCH PROCESS DATA

In many batch processes each batch run has the same duration and the variables follow closely some predetermined trajectories. However, in other cases the duration of the process varies from batch to batch. For example, a batch process where an exothermic chemical reaction occurs will vary in duration between summer and winter, because the cooling water temperature does not allow for as rapid heat removal in the summer. As a result, one cannot use as much initiator and catalyst and therefore the reaction takes longer. Another example is where polymerization reactions vary in duration owing to variation in the amount of impurities in the raw materials. In these situations the basic shapes of the time trajectories of each variable from batch to batch are similar, but their time duration varies. In order to analyze such data, it is necessary to account properly for the changing duration of the batches. This basically involves performing some form of realignment of the batch data prior to analysis.

Several very successful approaches have been used to align batch data. If a monotonically increasing or decreasing variable exists that always starts at a given level and ends at another specified level, then this 'indicator' variable may be useful for aligning the batches,[3] particularly if it is a good indicator of the 'maturity' or 'per cent completion' of each batch. To align the batches, this variable is simply used in place of time, and all other variables are plotted against it. Kourti *et al.*[22] analyzed data from an industrial semi-batch emulsion polymerization reactor where the batch durations varied by up to 20%. However, by using as an indicator variable the cumulative amount of a key monomer fed to the reactor over time (always starting at zero and ending with the same total amount), the batches were almost perfectly aligned. Neogi and Schlags[23] used the on-line measurement of a variable related to total conversion of the monomers to align batches from a polymerization process.

Once the batches are aligned, they can be analysed by any of the three-way methods by replacing the time direction with the indicator variable direction. For monitoring of new batches, data are simply collected and analysed at specified intervals of the indicator variable (see e.g. Reference 23). In this approach to alignment a measure of the 'maturity' or 'per cent completion' of any batch is provided by the percentage of its final value that has been attained by the indicator variable at the current point in time.

If no such indicator variable is available, then alternative approaches have been suggested. If one can define an end-point for any batch by the achievement of some event that is eventually reached in all batches, then this point can be used to define a 100% completion or 100% maturity point. For example, the batches could then be aligned using linear time adjustment by dividing each time point along the trajectory by the time at the 100% completion point. However, this only results in a linear compression or expansion of each batch, which is often inadequate for aligning batch trajectories (see e.g. Reference 24). Many batch runs occur in stages, and often it is only a few of those stages that proceed more slowly, while other stages may proceed at a normal pace or even more quickly. In this situation an approach based on ideas from the speech recognition literature, dynamic time warping (DTW), can be used for the non-linear alignment and synchronization of the batches.[24] In DTW each time point of the batch is shifted backward or forward in time in order to minimize the distance between the batch and a reference batch. DTW can also be used to predict the per cent completion of

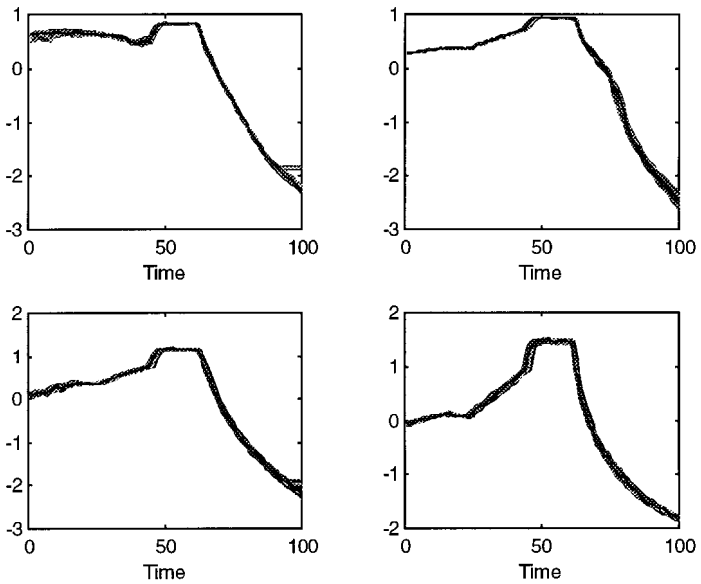

Figure 9. Four temperature profiles of 36 batches of DuPont data set

the batch (J. A. Westerhuis *et al.*, unpublished results). However, applying it in real time is more difficult than for the indicator variable approach.

Wold *et al.*[14] recently suggested another approach to predicting batch 'maturity' and aligning. In

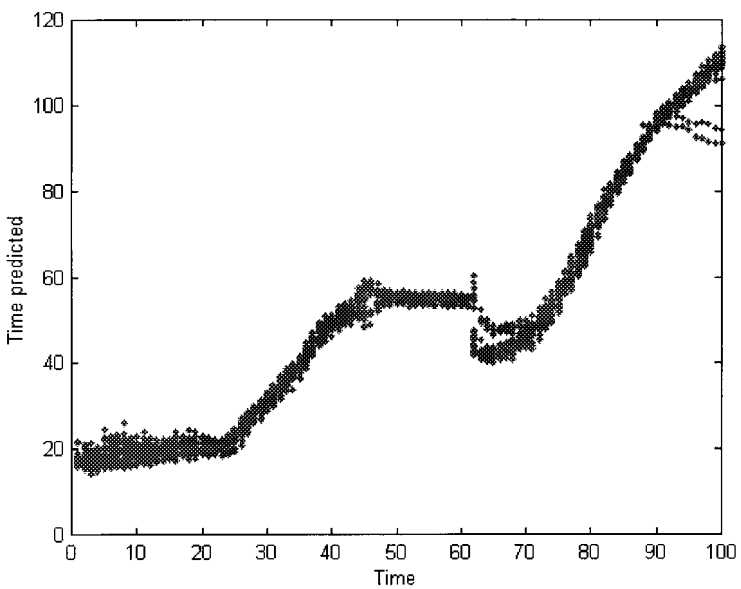

Figure 10. Predicted vs. real local batch time of a PLS model of four temperatures of 36 batches of DuPont data set

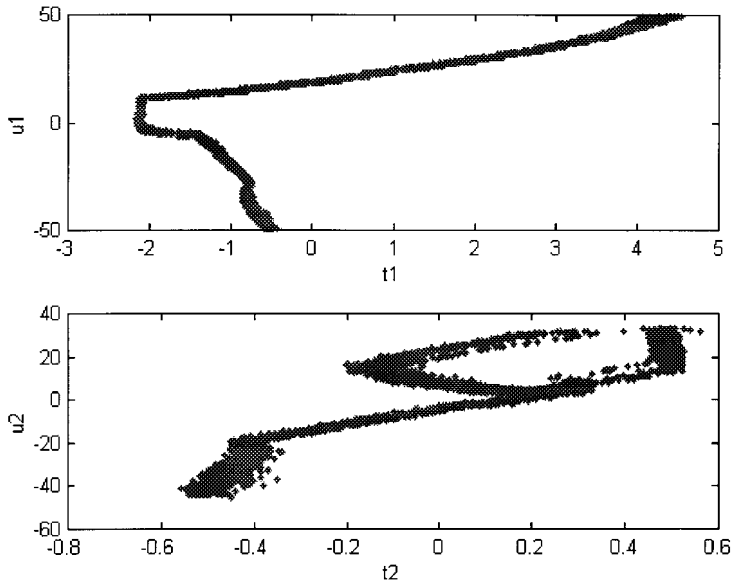

Figure 11. Non-linear relation between **t** and **u** scores of PLS model between process variables and local batch time of four temperatures of DuPont data set

this approach the three-way array $\underline{\mathbf{X}}$ is unfolded in the variable direction as matrix **A** in Figure 5. A PLS model is then developed between the unfolded matrix and the local batch time for each batch. This model is then used for new batches to predict a batch time, which is then used as an indicator of 'maturity' of the batch. This approach assumes that there is sufficient information in the trajectories to obtain a good prediction of batch time. This requires that there be some linear combination of the variables in each region of time that is either decreasing or increasing. An illustration of where one can have problems is shown in Figures 9–11. Suppose one has only the four variables from the DuPont data that are plotted in Figure 9. The PLS predictions of time for these batches as shown in Figure 10 are very poor. Note that during periods where all variables in Figure 9 are relatively constant, the time prediction is also constant and the predicted time can actually decrease as seen at around 60 min, because the settings of the process variables around this period are very similar to those at the earlier period around 40 min. An indication that for these data the method is having difficulties is revealed in the score plots in Figure 11, where extremely non-linear relationships between the latent variables for the process and time variable spaces are apparent.

In this section we have reviewed several approaches to aligning batch data. The simplest approach, and the one that has proven most useful to date with industrial data, is to replace time by an indicator variable. If such a variable does not exist, other approaches can be used, but some caution must be exercised in applying them.

5. CONCLUDING REMARKS

In this paper, several approaches to batch process data analysis that have appeared in the literature are discussed. PARAFAC, Tucker3 and MPCA models are compared for their abilities to model two previously published sets of data from batch polymerization processes.[1,3] Based on the results from these data and on a discussion of the nature of batch data and the suitability of the various methods for modeling these data, it is argued that MPCA is the preferred three-way modeling approach of batch

processes. In both batch data sets, MPCA was able to explain much more of both the fitted data and the cross-validated predictions and it did so with fewer latent variables in the important batch direction.

An important step in the preprocessing of batch process data is the unfolding and centering to remove the mean trajectories of the process variables. By removing the mean trajectories, a large part of the systematic variation in the data is removed and only the variation between the batches remains. This specific variation is useful for monitoring new batches because it detects if the systematic variation of the new batch is consistent with the variation in the good batches used to develop the model. If the mean trajectories are not removed from the data, many PCA components or PARAFAC components are necessary to describe the systematic variation between the batches. Proper scaling of the data is also important in batch process monitoring. Scaling should be chosen such that it improves the detection of faults.

If batches have unequal duration, it is of interest to align the data and to predict the per cent completion of the batch. The approach using an indicator variable to replace time is recommended if such a variable exists. Alternative approaches and potential problems have been discussed.

ACKNOWLEDGEMENTS

The calculations of the PARAFAC and Tucker models were performed using the *N*-way toolbox of Andersson and Bro.[18] We thank Rasmus Bro from the Department of Dairy and Food Science–Food Technology, Royal Veterinary and Agricultural University, Denmark, for his helpful comments and suggestions on this paper.

REFERENCES

1. P. Nomikos and J. F. MacGregor, 'Monitoring of batch processes using multiway principal component analysis', *AIChE J.* **40**, 1361–1375 (1994).
2. P. Nomikos and J. F. MacGregor, 'Multi-way partial least squares in monitoring batch processes', *Chemometrics Intell. Lab. Syst.* **30**, 97–108 (1995).
3. P. Nomikos and J. F. MacGregor, 'Multivariate SPC charts for monitoring batch processes', *Technometrics*, **37**, 41–59 (1995).
4. T. Kourti and J. F. MacGregor, 'Process analysis, monitoring and diagnosis, using multivariate projection methods', *Chemometrics Intell. Lab. Syst.* **28**, 3–21 (1995).
5. T. Kourti and J. F. MacGregor, 'Multivariate SPC methods for process and product monitoring', *J. Qual. Technol.* **28**, 409–428 (1996).
6. T. Kourti, P. Nomikos and J. F. MacGregor, 'Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS', *J. Process Control*, **5**, 277–284 (1995).
7. S. Wold, P. Geladi, K. Esbensen and J. Öhman, 'Multi-way principal components and PLS analysis', *J. Chemometrics*, **1**, 41–56 (1987).
8. A. K. Smilde and D. A. Doornbos, 'Three-way methods for the calibration of chromatographic systems: comparing PARAFAC and three-way PLS', *J. Chemometrics*, **5**, 345–360 (1991).
9. A. K. Smilde and D. A. Doornbos, 'Simple validatory tools for judging the predictive performance of PARAFAC and three-way PLS', *J. Chemometrics*, **6**, 11–28 (1992).
10. A. K. Smilde, 'Three-way analysis. Problems and prospects', *Chemometrics Intell. Lab. Syst.* **15**, 143–157 (1992).
11. R. Bro, 'PARAFAC. Tutorial and applications', *Chemometrics Intell. Lab. Syst.* **38**, 149–171 (1997).
12. P. Geladi, 'Analysis of multi-way (multi-mode) data', *Chemometrics Intell. Lab. Syst.* **7**, 11–30 (1989).
13. K. S. Dahl, M. J. Piovoso and K. A. Kosanovich, 'Translating third-order data analysis methods to chemical batch processes', *Chemometrics Intell. Lab. Syst.* **46**, 161–180 (1999).
14. S. Wold, N. Kettaneh, H. Friden and A. Holmberg, 'Modelling and diagnosis of batch processes and analogous kinetic experiments', *Chemometrics Intell. Lab. Syst.* **44**, 331–340 (1998).
15. B. M. Wise, 'A comparison of multiway principal components analysis, tri-linear decomposition and parallel

factor analysis for fault detection in a semiconductor etch process', presented at *Int. Chemometrics Research Meet.*, ICRM98, Veldhoven, 1998.
16. J. F. MacGregor and P. Nomikos, 'Monitoring batch processes', in *Batch Processing Systems Engineering*, Vol. 143 of *NATO ASI Series F*, Computer and Systems Sciences, ed. by G. V. Reklaitis, A. Y. Sunol, D. W. Rippin and O. Hortaçsu, pp. 241–258 (1992).
17. H. A. L. Kiers, 'Hierarchical relations among three way methods', *Psychometrica*, **56**, 449 (1991).
18. A. C. Andersson and R. Bro, 'The *N*-way toolbox for MATLAB', Royal Veterinary and Agricultural University, Denmark, http://newton.foodsci.kvl.dk/Matlab/nwaytoolbox.
19. R. Bro, 'Multi-way analysis in the food industry', *Doctoral Thesis*, Royal Veterinary and Agricultural University, Denmark (1998).
20. J. Pettersen, 'Application of statistical batch analysis to fermentation data', presented at *5th Nordic Workshop on Chemometrics (SUDMALT)*, Armentarola, Alta Badia, 1998.
21. P. Nomikos, 'Statistical process control of batch processes', *PhD Thesis*, McMaster University, Hamilton, Ontario (1995).
22. T. Kourti, J. Lee and J. F. MacGregor, 'Experiences with industrial applications of projection methods for multivariate statistical process control', *Comput. Chem. Eng.*, **20**, S745–S750 (1996).
23. D. Neogi and C. E. Schlags, 'Multivariate statistical analysis of an emulsion batch process', *Ind. Engng. Chem. Res.* **37**, 3971–3979 (1998).
24. A. Kassidas, J. F. MacGregor and P. A. Taylor, 'Synchronization of batch trajectories using dynamic time warping', *AIChE J.* **44**, 864–875 (1998).

