Research Article

Journal of CHEMOMETRICS

Received: 28 November 2006; Revised: 7 September 2007; Accepted: 29 November 2007 Published online in Wiley Interscience: 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1113

Bilinear modelling of batch processes. Part I: theoretical discussion

José Camacho[a]*, Jesús Picó[a] and Alberto Ferrer[b]

When studying the principal component analysis (PCA) or partial least squares (PLS) modelling of batch process data, one realizes that there is a wide range of approaches. In many cases, new modelling approaches are presented just because they work properly for a particular application, for example, on-line monitoring and a given number of processes. A clear understanding of why these approaches perform successfully and which are the advantages and disadvantages in front of the others is seldom supplied. Why does modelling after batch-wise unfolding capture changing dynamics? What are the consequences of variable-wise unfolding? Is there any best unfolding method? When should several models for a single process be used? In this paper, it is shown how these and other related questions can be answered by properly analyzing the dynamic covariance structures of the various approaches. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: principal component analysis; partial least squares; batch processes; unfolding methods

1. INTRODUCTION

Industrial batch processes produce a enormous amount of three-way data which is recorded for on-line treatment or posterior analysis. These are, nonetheless, difficult tasks due to the volume of the data and a low signal-to-noise ratio. One of the most widely used methods to extract the information from that kind of data are the projection to latent structures-based methods (PLS-based), like principal component analysis (PCA) and partial least squares (PLS). Since the duration of the processing of a batch may be variable in some processes, data have to be properly aligned—equalized—to apply most PLS-based methods. Several aligned methods can be found in the literature [1–6]. After the alignment, the data matrix of a batch process $\underline{\boldsymbol{X}}(I \times J \times K)$ contains the values of J variables at K sampling times in I batches.

Optionally, a number of initial conditions and quality measurements may be collected in a process. The L initial conditions measured for the I batches are arranged in matrix $\boldsymbol{Z}(I \times L)$. The M quality variables, collected at K_y sampling times, are stored in matrix $\underline{\boldsymbol{Y}}(I \times M \times K_y)$. Commonly, the quality variables are only measured at the end of the batch. Therefore, they may be arranged in a two-way matrix $\boldsymbol{Y}(I \times M)$.

The approaches for modelling batch processes with PLS-based methods can be roughly classified into two categories: The single model approach and the multi-model approach.

In the single model approach, a single PLS-based model is generated for the whole process. Although three-way modelling methods exist [7–10], the most commonly used procedure is to unfold the three-way matrix of data in two dimensions [11,12]. Afterwards, a bilinear model such as PCA or PLS can be fitted.

When using a single bilinear model, the direction of unfolding depends on the variability which is wanted to be modelled. According to Reference [13], for the treatment of batch process data, only batch-wise unfolding (1) and variable-wise unfolding (2) have interest

$$\underline{X}(I \times J \times K) \Rightarrow \mathbf{X}(I \times JK) \tag{1}$$

$$\underline{\mathbf{X}}(I \times J \times K) \Rightarrow \mathbf{X}(KI \times J) \tag{2}$$

where $\boldsymbol{X}$ is the unfolded two-way matrix of data. Batch-wise unfolding treats the data of a complete batch as an object (row) of the resulting two-way matrix of data (see Figure 1(a)). Variable-wise unfolding treats data collected each sampling time of a batch as an object (Figure 1(b)). Moreover, since data are usually centred and scaled to unit variance after the unfolding operation, additional differences between both unfolding methods are found because of the preprocessing.

The direction of unfolding also influences the design of an on-line application. If batch-wise unfolding is used [1], the measurements which are not available have to be imputed at every sampling time. The methods designed to impute missing data from a PCA and PLS model [14,15] can be used for this purpose. If variable-wise unfolding is used no imputation is necessary, but the dynamics of the process in the form of auto-covariances and lagged cross-covariances among variables are not captured [13]. Some authors have proposed to combine both unfolding methods. In Reference [3], three monitoring levels are generated, where only the first two are used for on-line monitoring. The level 1 consists of a variable-wise model. Thus, it does not take into account the auto-covariances and lagged cross-covariances. The level 2 uses the scores obtained from the previous model to generate the monitoring charts. This is carried

* *Departamento de Ingeniería de Sistemas y Automática, Universidad Politécnica de Valencia, Valencia, Spain.*
E-mail: jcamacho@isa.upv.es

a *J. Camacho, J. Picó*
Departamento de Ingeniería de Sistemas y Automática, Universidad Politécnica de Valencia, Valencia, Spain

b *A. Ferrer*
Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia, Valencia, Spain

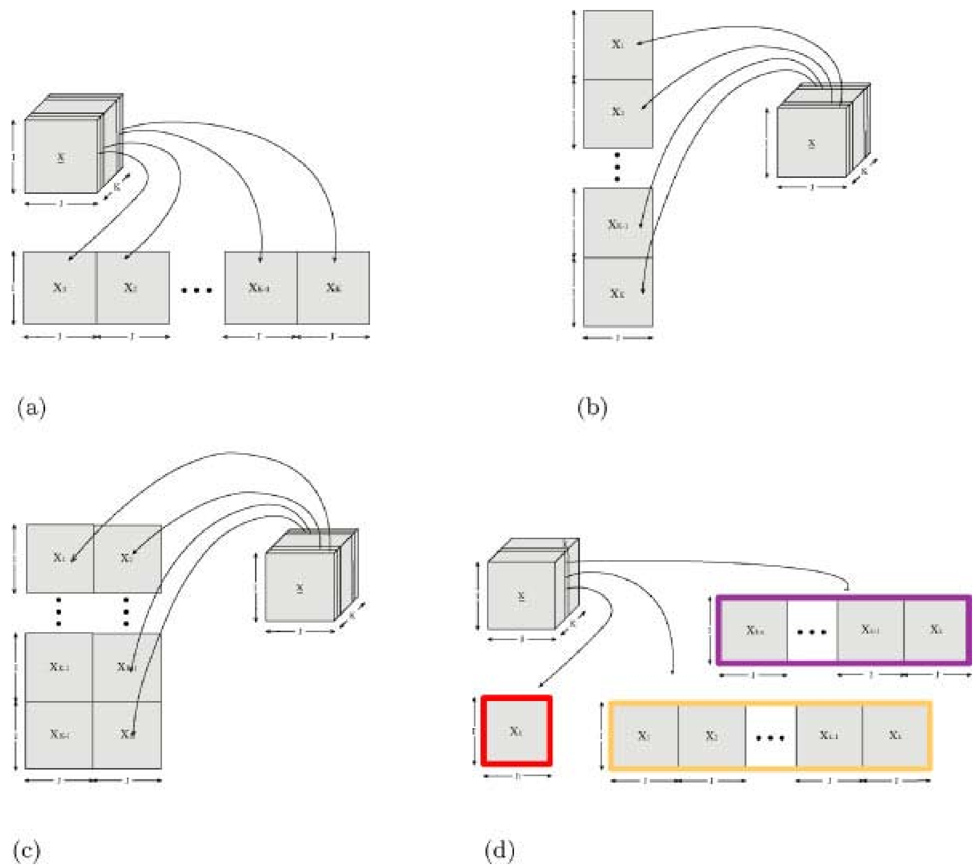

Figure 1. Different arrangements of three-way batch data in two-way form. (a) batch-wise unfolding, (b) variable-wise unfolding, (c) batch dynamic unfolding with 1 LMV, (d) *K*-models approaches. In (d), from left to right, the data used to fit a local model, an evolving model and a moving window model for sampling time *k* are represented. This Figure is available in colour online at www.interscience.wiley.com/journal/cem

out by rearranging the scores as in batch-wise unfolding. The authors also conceive the possibility of generating a PCA model from these rearranged data. By doing so, auto-covariances and lagged cross-covariances of the scores of level 1 are captured in the level 2 [3]. Alternatively, Reference [16] proposes to take advantage of both batch-wise and variable-wise unfolding procedures by generating two models from the original data, one after each unfolding.

Batch dynamic PCA (BDPCA) and batch dynamic PLS (BDPLS) [17] are another single-model approaches equivalent to the variable-wise unfolding with the addition of lagged measurement vectors (LMVs) as variables [18]. An example of this unfolding with 1 LMV is shown in Figure 1(c). The batch dynamic unfolding can be seen as a generalization of the traditional unfolding procedures: if no LMV is added, the resulting matrix is the same as the one after variable-wise unfolding; if all possible LMVs are added, the resulting matrix is the same as the one after batch-wise unfolding.

The multi-model approach is traditionally based on the generation of a bilinear model for every sampling time of the batch duration [19–21]. This is referred here as the *K*-models approach (Figure 1(d)). These models do not need to impute measurements and are able to capture the dynamics by including LMVs as variables, but the price to pay is the generation of that high number of models. Again, several proposals can be found in the literature, which mainly differ in the data used to generate the sub-models. If each sub-model includes only the data of a sampling time, then it is called a local model. If each sub-model incorporates the measurements from the beginning of the batch to a sampling time, then it is called an evolving model [20]. Hierarchical models [19] follow the latter method but giving different weight to the current measurements of the variables. This is done by calculating the scores in a hierarchical fashion. Finally, if only the immediate part of the past measurements is included in each sub-model together with the current measurements, the procedure is called moving window PCA (MWPCA) [21] or moving window PLS (MWPLS).

One of the most appreciated benefits of using PLS-based techniques is that, by conveniently studying the fitted model, the process understanding can be improved [22]. Although the calibration of a sub-model for each sampling time can be advantageous for the on-line application, such a number of sub-models can be complex to handle and difficult to interpret. Recently, the use of a reduced number of sub-models, each one representing a certain period of the process, has been studied. In References [16] and [23], authors propose the use of a sub-model for every processing unit of the process. Also, several sub-models can be defined for a single processing unit if necessary. Reference [24] proposes a clustering algorithm for the automatic identification of a number of sub-models for on-line monitoring. The same objective is pursued in Reference [18], where the sub-models are identified using the multi-phase PCA algorithm [25].

In this paper, a theoretical discussion regarding the differences among the principal modelling approaches is carried out by studying the structure of the resulting covariance matrix. This discussion is limited to bilinear and non-hierarchical PLS-based models. In Section 2, some useful notation is introduced. In Section 3, the use of the covariance matrix for the analysis of bilinear methods is briefly discussed. In Section 4, the single model approach is studied. In Section 5, the multi-model approach based in the design of a single PLS-based model for every sampling time, that is, the *K*-models approach, is analysed. In Section 6, the multi-phase modelling approach is presented. Finally, Section 7 gives some conclusions.

2. SOME NOTATION

To apply a bilinear modelling method, data in matrix $\underline{\boldsymbol{X}}(I \times J \times K)$ have to be rearranged in two dimensions. This can be achieved by unfolding $\underline{\boldsymbol{X}}$, by dividing $\underline{\boldsymbol{X}}$ in a number of two-way matrices or by a combination of both.

2.1. Unfolding

The batch dynamic unfolding can be expressed as

$$\boldsymbol{X} = \text{unfold}\,(\underline{\boldsymbol{X}}, n) \equiv \underline{\boldsymbol{X}}^{(n)} \tag{3}$$

where n stands for the number of LMVs included as variables and

$$n = \{k - 1 : k \in \{1, 2, \ldots, K\}\} \tag{4}$$

Therefore, the batch-wise unfolding is

$$\boldsymbol{X} = \underline{\boldsymbol{X}}^{(K-1)} \tag{5}$$

and the variable-wise unfolding

$$\boldsymbol{X} = \underline{\boldsymbol{X}}^{(0)} \tag{6}$$

2.2. Number of sub-matrices

Let $\underline{\boldsymbol{X}}_{k_i:k_e}$ contain the data of $\underline{\boldsymbol{X}}$ from sampling time k_i to k_e. One way to arrange $\underline{\boldsymbol{X}}$ in two-way is to divide the data in K local sub-matrices

$$\boldsymbol{X} = \{\boldsymbol{X}_k : k = 1, \ldots, K\} \tag{7}$$

where $\boldsymbol{X}_k \equiv \underline{\boldsymbol{X}}_k^{(0)}$.

Combining the unfolding with the division in several sub-matrices, many other approaches can be specified. For instance, the Evolving approach

$$\boldsymbol{X} = \left\{\underline{\boldsymbol{X}}_{1:k}^{(k-1)} : k = 1, \ldots, K\right\} \tag{8}$$

3. STUDY OF THE COVARIANCE MATRIX

Strictly speaking, the analysis carried out in this paper is focused on matrix $\boldsymbol{X}^T \cdot \boldsymbol{X}$. This matrix is the same, but a constant factor, to the covariance matrix of $\boldsymbol{X}$ when it is centred. By inspecting $\boldsymbol{X}^T \cdot \boldsymbol{X}$ the dynamic and static relationships among process variables built in a model can be observed.

The analysis of $\boldsymbol{X}^T \cdot \boldsymbol{X}$ is valid to understand the dynamics built in PCA models (and so it is for principal components regression (PCR) models). In the case of PLS, let us assume the matrix $\boldsymbol{X}^T \cdot \boldsymbol{Y} \cdot \boldsymbol{Y}^T \cdot \boldsymbol{X}$ is used to fit the models. Making the change of variable $\boldsymbol{C} = \boldsymbol{Y}^T \cdot \boldsymbol{X}$ it can be seen that PLS is performed over $\boldsymbol{C}^T \cdot \boldsymbol{C}$. Therefore, the conclusions drawn from $\boldsymbol{X}^T \cdot \boldsymbol{X}$ are also valid for PLS but with the difference that in the latter the relationships are not among original process variables, but among covariance values between process and quality variables.

4. SINGLE MODEL APPROACH

Two traditional methods have been proposed for the end-of-batch and on-line monitoring of batch processes: the approach of Nomikos and MacGregor [1,11] and the approach of Wold *et al*. [3]. The principal difference between these methods is found in the modelling for on-line monitoring. In Reference [1], the batch-wise unfolding (Figure 1(a)) procedure is used. In Reference [3], the variable-wise unfolding (Figure 1(b)) is used as first step to create the monitoring system. In both cases, the unfolded data are auto-scaled, that is, each variable of the unfolded matrix is centred and scaled to unit variance. This implies that the bilinear models in both approaches are modelling different data, since for the first approach the average trajectory of the process variables is subtracted whereas for the second one it is not. Therefore, the influence of the unfolding direction and of the preprocessing method in the differences found between both proposals is not clearly distinguishable.

Although the implications of the unfolding direction in a on-line system have been pointed out by several authors [3,13,26] and introduced in the first section of this paper, here this issue is more deeply revisited focusing on the structure of the resulting covariance matrices and with an independent point of view from the preprocessing mechanism. Furthermore, these implications do not seem to be very well understood by part of the research community yet. For instance, some authors believe that by batch-wise unfolding, the resulting model does not include the dynamics of the variation around the average trajectory [17]. It will be shown that this conclusion is wrong.

In the following, let imagine the matrix $\underline{X}$ corresponds to the data collected from a batch process after the subtraction of the average trajectory.[†]

4.1. Batch-wise unfolding

In Figure 2, the resulting $JK \times JK$ covariance matrix after the batch-wise unfolding of the data, that is, $\underline{\boldsymbol{X}}^{(K-1)}$ (Figure 1(a)) is shown. It can be seen that the relationships among all the variables at different sampling times are modelled by the PCA performed with this matrix. These relationships include the variances and instantaneous cross-covariances of the variables at every sampling time—represented by the square sub-matrices $\{VC_1 \cdots VC_K\}$ in the diagonal—along with the auto-covariances and lagged cross-covariances of the variables—relationships outside these sub-matrices. Sub-matrices $V_{k,k+d}$ for $k = 1, \ldots, K - d$ contain the auto-covariances and cross-covariances of order d. In Figure 2, only sub-matrices $V_{k,k+1}$ are drawn for the shake of simplicity.

The dynamics of the variation around the average trajectory are represented by the relationships in sub-matrices $V_{k,k+d}$. It can be concluded that, unlike some authors have claimed, batch-wise unfolded models do incorporate the linear dynamics (around the average trajectory) of the process. Furthermore, these models are able to capture changing relationships among variables, since sub-matrices VC_k and $V_{k,k+d}$ for two different values of k—say 1

[†] The average trajectory is subtracted so that the resulting two-way matrix after both batch-wise and variable-wise unfolding is centred and the same data are modelled.

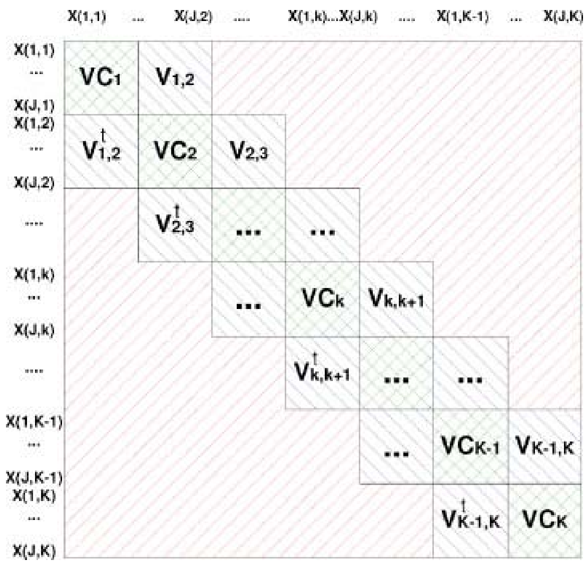

Figure 2. Covariance matrix after the subtraction of the average trajectory and batch-wise unfolding. This Figure is available in colour online at www.interscience.wiley.com/journal/cem

and 10—are placed in different parts of the covariance matrix, and so treated independently.

4.2. Variable-wise unfolding

The $J \times J$ covariance matrix obtained after variable-wise unfolding, that is, $\underline{\boldsymbol{X}}^{(0)}$ (Figure 1(b)) is equivalent, but a constant factor equal to $\frac{1}{K}$, to the sum of the matrices $\{VC_1 \cdots VC_K\}$ (see Figure 3). This has two consequences: first, variable-wise models only incorporate the variances and instantaneous cross-covariances of the variables and not the dynamics; and second, each relationship between a pair of variables in the covariance matrix is an average of their relationship throughout the batch duration. Thus, this modelling strategy is only valid when the correlation structure of a process is more or less constant. On the other hand, the number of batches required to generate a model in variable-wise unfolding is lower than in batch-wise unfolding. After the former unfolding, for every batch, a number of objects equal to the length of the batch is available. In batch-wise unfolding, a batch is a single object.

4.3. Batch dynamic unfolding

The batch dynamic unfolding is equivalent to the variable-wise unfolding with LMVs addition [18]. Figure 4 shows the corresponding covariance matrix after data have been unfolded

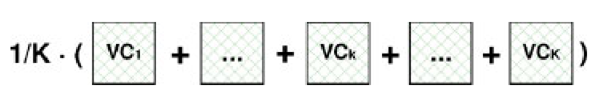

Figure 3. Covariance matrix after the subtraction of the average trajectory and variable-wise unfolding.

following the example in Figure 1(c), $\underline{\boldsymbol{X}}^{(1)}$. From this Figure it is easy to understand the effect of the addition of LMVs. Dynamics of order d are built in the model by matrices $V_{k,k+d}$. The more LMVs added, the more dynamic information included in the model. As stated in Reference [25], depending on the nature of the process, part of the dynamic information may be negligible and it could be advantageous not to incorporate it to the model. Thus, identifying the optimum number of LMVs from the data of a process is, *a priori*, a more powerful modelling approach than using an extreme case, like in batch-wise or variable-wise unfolding.

It should be pointed out that when the number of LMVs is low, the resulting batch dynamic model is very close to the traditional way of modelling the dynamics with autoregressive models. Auto-correlations and cross-correlations are obtained independently of the precise sampling time. This model structure is very useful for regulation and can be used to relax the necessity of alignment of the batches [5]. Nonetheless, as it happens for variable-wise unfolded models, modelling with a reduced number of LMVs assumes a constant correlation structure during the batch. The more LMVs included in the model, the less the number of objects -rows- in the unfolded matrix—$\underline{\boldsymbol{X}}^{(n)}$ has $I(K-n)$ rows. Therefore, by adding LMVs, the interval where the correlation structure is imposed to be constant is reduced and the model is more capable to capture changes in the dynamics. At the same time, the dynamics built in the model are more dependent on the precise sampling time. A batch dynamic model of the form $\underline{\boldsymbol{X}}^{(n)}$ assumes the dynamics of the process can be considered time-invariant for a $K-n$ sampling times period.

Although batch dynamic unfolding provides the possibility to capture time-varying correlation structures by adding a sufficient number of LMVs, this may lead to over-parameterized models. Take the extreme example of a process in which time-varying but only static relationships among variables exist. This process can be modelled by batch-wise unfolding the data and applying a bilinear PLS-based method. Then, time-varying static relationships will be effectively modelled by matrices VC_k in Figure 2. Nonetheless, the rest of the covariance matrix will be just noise, since there are not significant dynamic relationships.

5. *K*-MODELS APPROACH

The *K*-models approach (see Figure 1(d)) is based on generating a single sub-model for every sampling time. Local models [20] are the simplest example of this approach, where the sub-model associated to a sampling time is computed from the data collected at that sampling time alone. Data are split according to Equation (7). Nonetheless, the inclusion of the dynamics of the process can be crucial for a good modelling performance. A straightforward alternative to local modelling is evolving modelling [20], where all the possible LMVs are included in a sub-model as additional variables (8).

More sophisticated approaches can be used. Not all the lagged information may be of interest or at least be given the same importance in every sub-model. Uniformly weighted moving

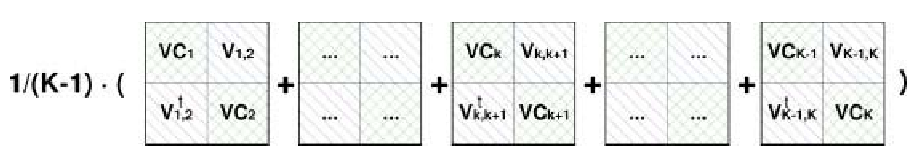

Figure 4. Covariance matrix after the subtraction of the average trajectory and batch dynamic unfolding according to Figure 1(c).

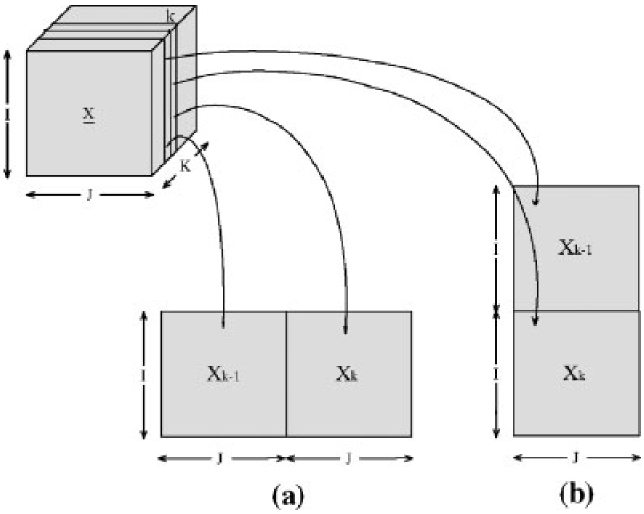

Figure 5. Addition of 1 LMV to the local matrix of data at sampling time k: as additional variables, columns (a) and as additional objects, rows (b).

window (UWMW) or Exponentially weighted evolving window (EWEW) models may be defined in order to better adjust the multi-model to the nature of the process. UWMW modelling uses the current data along with those of the immediate n_k LMVs to generate the current sub-model

$$\boldsymbol{X} = \{\underline{\boldsymbol{X}}_{k-n_k:k} : k = 1, \ldots, K\} \tag{9}$$

n_k is *the size of the window*, a calibration parameter of the UWMW model.

An EWEW sub-model includes all the lagged measurements to the current sampling time:

$$\boldsymbol{X} = \{\underline{\boldsymbol{X}}_{1:k} \odot \underline{\boldsymbol{W}}_{1:k} : k = 1, \ldots, K\} \tag{10}$$

where $\underline{\boldsymbol{W}}_{1:k}$ is a weighting matrix and $\odot$ stands for the Hadamard (element to element) product. $\underline{\boldsymbol{W}}_{1:k}$ is constructed according to an exponentially decreasing value, the *forgetting factor* $\lambda_k \in [0, 1]$, so that each measurement losses importance as the process advances. The weight of the measurement-vector collected at time $k - d$, for the generation of the sub-model at time k, is $(\lambda_k)^d$, being the weight of the current measurements always $(\lambda_k)^0 = 1$.

The parameters n_k and λ_k in UWMW and EWEW models are calibrated independently for every single sub-model—that is why the subscript k is used—for instance by using cross-validation.

Additionally, there are two ways to carry out the inclusion of lagged measurements: as additional variables—columns—or as additional objects—rows—of the unfolded matrix. In Figure 5, these two ways are shown for the addition of 1 LMV to the current measurement vector at sampling time k.

In Figure 6, the covariance matrices of an UWMW model for the inclusion of the lagged measurements as variables or objects, are shown. Both arrangements can be respectively noted as

$$\boldsymbol{X} = \{\underline{\boldsymbol{X}}^{(n_k)}_{k-n_k:k} : k = 1, \ldots, K\} \tag{11}$$

and

$$\boldsymbol{X} = \{\underline{\boldsymbol{X}}^{(0)}_{k-n_k:k} : k = 1, \ldots, K\} \tag{12}$$

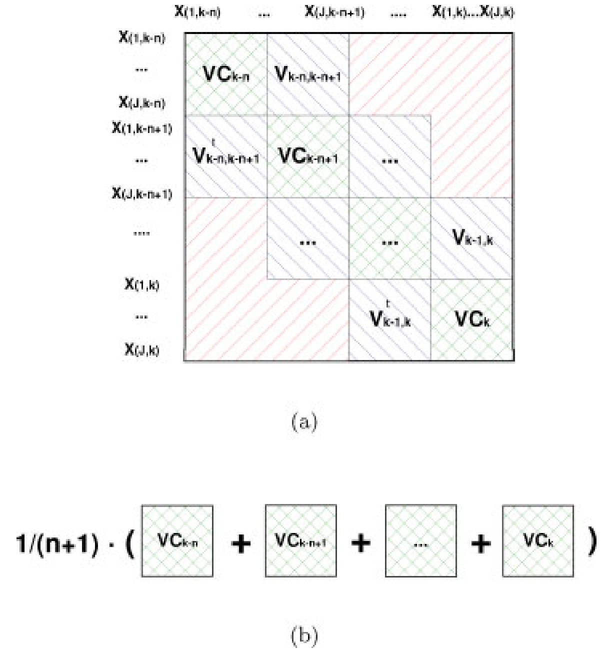

Figure 6. Covariance matrices of UWMW models: LMVs added as variables, columns (a) or as objects, rows (b). n is the size of the window. This Figure is available in colour online at www.interscience.wiley.com/journal/cem

In Figure 7, the same is shown for an EWEW model. Also, both arrangement approaches can be noted as:

$$\boldsymbol{X} = \left\{\underline{\boldsymbol{X}}^{(k-1)}_{1:k} \odot \underline{\boldsymbol{W}}^{(k-1)}_{1:k} : k = 1, \ldots, K\right\} \tag{13}$$

and

$$\boldsymbol{X} = \{\underline{\boldsymbol{X}}^{(0)}_{1:k} \odot \underline{\boldsymbol{W}}^{(0)}_{1:k} : k = 1, \ldots, K\} \tag{14}$$

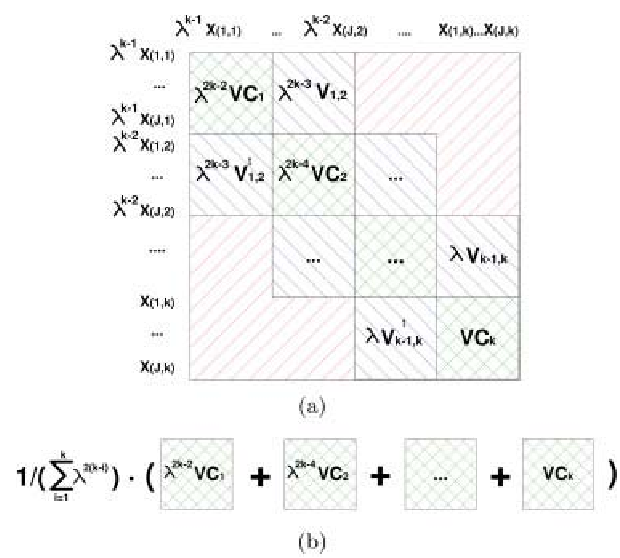

Figure 7. Covariance matrices of EWEW models: LMVs added as variables, columns (a) or as objects, rows (b). λ is the forgetting factor. This Figure is available in colour online at www.interscience.wiley.com/journal/cem

The two modelling approaches have different properties. The EWEW models take into account all the past information. Thus, they handle a bigger amount of data than UWMW models and so they are more computationally demanding. Nonetheless, a recursive PLS algorithm [27,28] may be applied to reduce that complexity. Both UWMW and EWEW coincides for $\lambda_k = 0$ and $n_k = 0$, values for which they become local models, and for $\lambda_k = 1$ and $n_k = k - 1$, for which they become evolving models when the LMVs are included as variables. Therefore, EWEW and UWMW strategies can be seen as a general parametrization where evolving and local models are extreme cases.

Comparing the covariance matrices of Figures 6(a) and 7(a) with those of Figures 6(b) and 7(b), it should be stressed that the inclusion of LMVs as variables is more similar to the batch-wise unfolding, whereas the effect of this inclusion as additional objects resembles the case of variable-wise unfolding. A direct conclusion might be extracted: the dynamic information contained in the auto-covariances and lagged cross-covariances is not captured with the inclusion of lagged objects. Notice that this information is contained in the sub-matrices of the general form $V_{t,t+d}$, which are only present in the covariance matrices of Figures 6(a) and 7(a). Since the dynamics are important in the modelling, a poor modelling performance of the approaches which do not include lagged variables is expected.

Finally, in Reference [19] a hierarchical *K*-models approach is presented. The basis of this approach is to combine the past and local information with an adaptive hierarchical PCA model. The model for every sampling time is obtained from the high level loadings of the immediate past sampling time model and the measurements collected at the current sampling time. The analysis of this approach from the covariance matrix is not trivial and it is not performed here. Nonetheless, in the companion paper, two extensions of this approach to PLS are compared with the other models studied here and differences are highlighted.

6. MULTI-PHASE APPROACH

The multi-phase approach is proposed to overcome the shortcomings of the single model approach with a reduced number of sub-models. First, batch-wise models may not perform adequately when several periods of the batch process are nonlinearly related or independent [25]. This is discussed in Section 6.1. Second, as commented before, variable-wise models cannot capture changing correlation structures and process dynamics. This is further studied in Section 6.2. Finally, batch dynamic models can be affected by both nonlinearity or independence and changes in the correlation structure, as it will be explained in Section 6.3. In the three cases, the model can be improved by modelling independently periods of the batch: the phases.

Let us define a phase as a segment of the batch duration which is well approximated by a single linear model. From this definition it should be stressed that the optimum division in phases of a process depends on the structure of the linear models used. Using batch-wise models, the phases are periods where the variables are linearly related. Using variable-wise models, the phases are periods with a constant correlation structure. Finally, for batch dynamic models, the phases are periods with constant dynamics and where the current and immediate preceding *n* LMVs are linearly related.

The multi-phase modelling approach provides more flexible models than the single model and *K*-models approaches. Nonetheless, we are aware that it has also its limitations. First, as it is well known, the use of local linear models for modelling nonlinearity does not take into account part of the multivariate nature of the data, that is, when approximating nonlinearity with local models, part of the information is lost. Second, the changes in the correlation structure of a batch may be smooth. In that case, smooth transitions between sub-models may be more adequate than a crisp division in phases. Although smooth transitions can be achieved with the currently available modelling techniques, for instance by using the Fuzzy theory or the Mixtures of Probabilistic PCA [29], it should be noted that these methodologies may complicate the use and easy interpretation of the models. The use of fuzzy logic for multi-phase modelling was originally proposed in Reference [25] and has been recently investigated in Reference [30].

6.1. Phases in batch-wise data

When data include independent or nonlinearly related groups of variables, a different model should be designed for each of these groups. This is discussed in detail in Reference [25]. Nonlinear relationships and independence between two variables are examples of relationships poorly approximated with a linear model. The variables so related show low correlation values. In batch processes, two observations of the vector of process variables commonly show less correlation as the time (grade of completion of the batch) distance between them grows. In Figure 8, the correlation map of the batch-wise unfolded data of two variables, collected from a polymerization process, is shown. This is a well-known real data set of 50 batches provided by Dupont Co. For more details of the process or the data set, see [22]. The data is aligned to a batch length of 116 sampling times. The result of the batch-wise unfolding is a bi-dimensional matrix with 50 observations of 232 variables, ordered as depicted in Figure 1(a).

The Figure shows a diagonal of high correlation. The dark rectangles of the graphic represent the segments of the batch where the variables are highly correlated with their preceding values. The correlation gets negligible as the time distance grows. The formation of rectangles in the correlation map means that

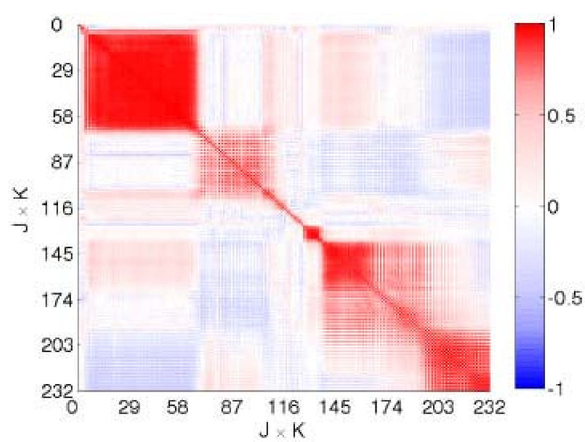

Figure 8. Correlation of two variables measured 116 times (batch duration) after batch-wise unfolding and preprocessing (auto-scaling). Nylon 6′6 cultivation process. This Figure is available in colour online at www.interscience.wiley.com/journal/cem

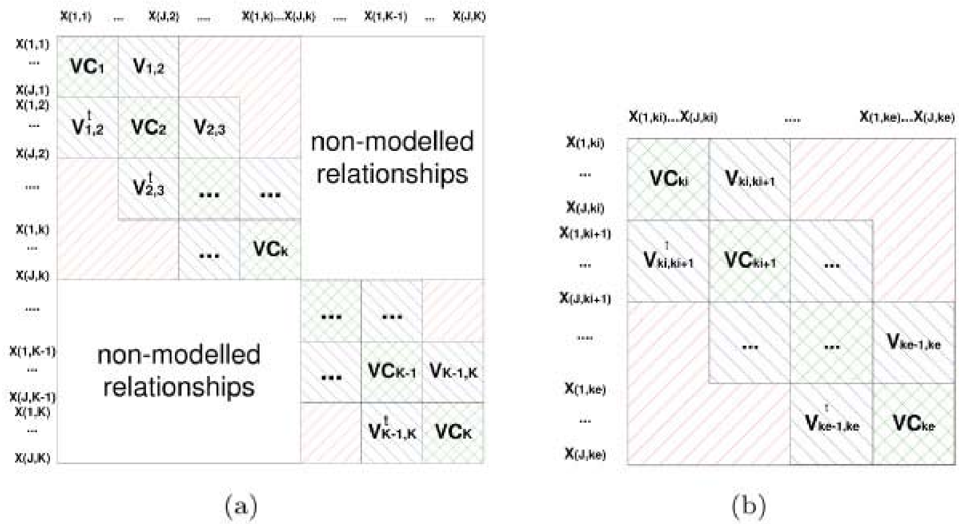

Figure 9. (a) Modelled and non modelled relationships after a division of a batch-wise model. (b) Covariance matrix of a sub-model in the multi-phase approach for batch-wise data. This Figure is available in colour online at www.interscience.wiley.com/journal/cem

the auto-correlation and lagged cross-correlation is dependent on the phase of the batch and that the modelling with linear models should be performed independently for different phases. For instance, the dark rectangle in the up-left corner shows that both process variables are highly correlated between each other and in time for approximately the first 30 sampling times—in the figure, from 1 to 60.

In Figure 9(a), the effect of the division in phases of batch-wise data in the resulting covariance matrix is shown. Part of the relationships are simply not modelled, but notice that this part may show low correlation values (as in the light areas of Figure 8). The resulting covariance matrix of the sub-model of a phase from sampling time k_i to k_e is presented in Figure 9(b).

Following the notation presented, a general multi-phase partition from batch-wise unfolded data is represented by

$$\mathbf{X} = \left\{ \underline{\mathbf{X}}_{k_{il}:k_{el}}^{(k_{el}-k_{il})} : l = 1, \ldots, L \right\}$$
$$s.t.\ k_{il} \leq k_{el}, k_{i(l+1)} = k_{el} + 1, k_{i1} = 1, k_{eL} = K \qquad (15)$$

It should be stressed that this approach is a generalization of the batch-wise unfolding. When properly calibrated, multi-phase models following (15) outperform batch-wise models in terms of prediction error [25] and are specially suited for off-line batch process monitoring [31,32].

6.2. Phases in variable-wise data

The variable-wise models have two principal drawbacks: they do not model dynamic information and impose a constant correlation structure in the data [33]. The first drawback can be overcome by adding LMVs, yielding batch dynamic models. The second can be overcome by using several sub-models.‡

The problem of modelling changing correlation structures is the same problem when the objects form clusters in the variables-space. The presence of clusters means that the objects of a cluster

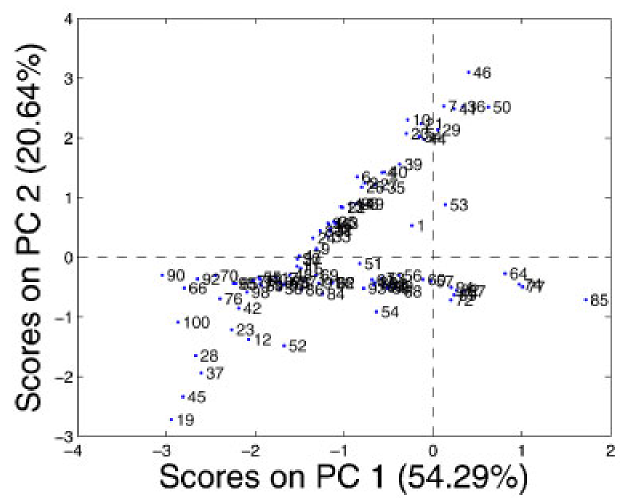

Figure 10. Scores of a batch for the 2 first PCs of a variable-wise model from data of the *Saccharomyces cerevisiae* cultivation process.

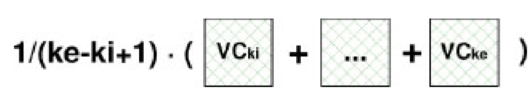

Figure 11. Covariance matrix of a sub-model in the multi-phase approach for variable-wise data.

represent a different reality than the objects of another cluster. This is discussed in Reference [32].

The first-principles model of the *Saccharomyces cerevisiae* cultivation process presented in Reference [34] was used to generate a data set of 30 batches in Reference [18]. Using these data, in Figure 10, the scores of a batch from a 2 PCs variable-wise model are shown. It can be seen that the distribution of the scores of the first 50 sampling times (approximately) is very different to that of the last 50 sampling times. These two groups of sampling times should be modelled separately.

The resulting covariance matrix of the sub-model of a phase from sampling time k_i to k_e is presented in Figure 11. The local

‡ Also, by adding a sufficient number of LMVs, changing correlation structures are effectively modelled at the expense of over-parametrization.

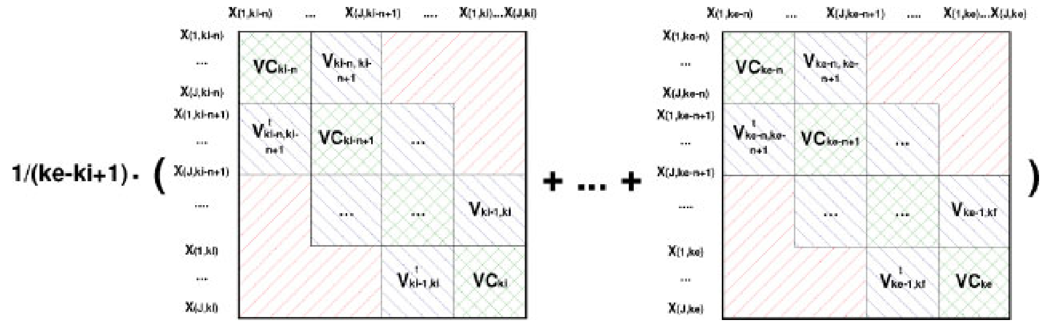

Figure 12. Covariance matrix of a sub-model in the multi-phase approach for batch dynamic data.

models of the single sampling times are averaged as in variable-wise unfolding.

A general multi-phase partition from variable-wise unfolded data is represented by

$$\boldsymbol{X} = \left\{\underline{\boldsymbol{X}}^{(0)}_{k_{il}:k_{el}} : l = 1, \ldots, L\right\}$$
$$s.t.\ k_{il} \leq k_{el}, k_{i(l+1)} = k_{el} + 1, k_{i1} = 1, k_{eL} = K \quad (16)$$

6.3. Phases in batch dynamic data

The resulting covariance matrix of the sub-model of a phase from sampling time k_i to k_e, for batch dynamic data, is presented in Figure 12. The covariance matrices of the UWMW models of the single sampling times—from k_i to k_e—are averaged.

The multi-phase models based on batch dynamic data use several sub-models for capturing changing correlation structures. The auto-covariances and lagged cross-covariances are modelled by including LMVs as variables. The size of the window n depends on the amount of past information needed. Each sub-model can have a different optimum n value. By properly calibrating this parameter, problems with nonlinear relationships or independence are also avoided.

A general multi-phase partition from batch dynamic unfolded data is represented by

$$\boldsymbol{X} = \left\{\underline{\boldsymbol{X}}^{(n_l)}_{\max(k_{il}-n_l,1):k_{el}} : l = 1, \ldots, L\right\} \quad (17)$$

with

$$n_l = \{k - 1 : k \in \{1, 2, \ldots, k_{el}\}\}$$
$$s.t.\ k_{il} \leq k_{el}, k_{i(l+1)} = k_{el} + 1, k_{i1} = 1, k_{eL} = K \quad (18)$$

Notice that the sub-models are not completely disjoint, since n_l measurement vectors are used to fit both sub-models $l-1$ and l.

The matrix of Figure 11 is a particular case (for $n = 0$) of that of Figure 12. Nonetheless, a multi-phase partition of batch-wise data (Figure 9) cannot be seen as a particular case of the multi-phase partition of batch dynamic data. This is a consequence of the definition in Equations (17) and (18). In this definition, the only possibility of obtaining disjoint sub-models is for variable-wise unfolding, that is, $n_l = 0$ for $l \in \{1, 2, \ldots, L\}$. Therefore, the multi-phase partition of batch-wise data in Equation (15), which also yields disjoint sub-models, is not contemplated in Equations (17) and (18), except for the case of a single phase batch-wise model.

The definition of Equations (17) and (18) is a general parametrization which includes single phase batch-wise models, variable-wise models, batch dynamic models and UWMW models, being more flexible than any of them. Although here UWMW models have been used to model the phases, an alternative approach is to use EWEW models.

The multi-phase models of batch dynamic data are specially suited for their use on-line [18,35].

7. CONCLUSIONS

This is the first paper of a series of two. In this paper, a theoretical discussion regarding the differences among several approaches for modelling batch process data based on the structure of the resulting covariance matrices is presented. This discussion is restricted to models based on the PLS-based, in particular to bilinear models like PCA and partial least squares (PLS). In the companion paper [35], an experimental comparative based on PLS is performed.

Approaches for modelling batch data using bilinear PLS-based models can be classified in two main groups: single model and multi-model approaches. The latter, in turn, can be divided in K-models approaches and multi-phase approaches. The single model approaches use the same model for all the batch duration. The K-models approaches fit a single model per sampling time. Finally, the multi-phase approaches fit a different model for certain periods (phases) of the batch, given a criterion for the identification of the periods which should be modelled separately.

In single model approaches, the three-way matrix of data is unfolded into a two-way matrix and then a bilinear modelling method is applied. The most well-known unfolding directions for batch process data, that is, batch-wise unfolding and variable-wise unfolding, can be seen as special cases of a general unfolding method: the batch dynamic unfolding. In the batch dynamic unfolding, a number of LMVs are added to the current measurement vector to form the object—row—of the unfolded two-way matrix. If no LMV is added, the resulting matrix is the same yielded with variable-wise unfolding. If all possible LMVs are added, then the unfolded matrix is the same yielded with batch-wise unfolding. The advantage of the batch dynamic unfolding is that the number of LMVs of the model can be optimized for a particular process. This is, *a priori*, a more powerful modelling approach than using an extreme case, like in batch-wise or variable-wise unfolding.

As it can be seen from the structure of the covariance matrices, there are two different reasons for the addition of LMVs in the single model approach: the modelling of the dynamics of a certain order and/or the modelling of changing correlation structures, understanding correlation structure as the way variables are related with each other and in time. The more LMVs added, the more dynamic information is included in the model. Therefore, variable-wise models do not capture the dynamics of the process whereas batch-wise models do. Moreover, modelling with a reduced number of LMVs assumes a constant correlation structure during the batch. The more LMVs included in the model, the shorter the interval where the correlation structure is imposed to be constant and the more capable the model is to capture changes in the dynamics. At the same time, the dynamics built in the model are more dependent on the precise sampling time. On the other hand, for the sake of parsimony of the resulting model,[§] the less the number of variables in the unfolded matrix the better. Also, the inclusion of independent or nonlinearly related variables in the same object is not recommendable for bilinear models. As a conclusion, the number of LMVs in a single model should be kept as low as possible, but being high enough to capture the—possibly changing—dynamics of the process. This compromising solution may tend to yield over-parameterized models.

The *K*-models approach is based on the calibration of one bilinear model per sampling time. By properly setting the number of LMVs in the models, this approach is able to capture the changing dynamics and avoids problems with nonlinearity and independence between different phases of the process. Nonetheless, its principal drawback is the high number of sub-models, which makes the calibration and interpretation challenging. This is also the less parsimonious approach and so a larger data set is needed for a proper calibration.

Multi-phase models overcome the limitations of single-model approaches with a reduced number of sub-models. In particular, the multi-phase modelling approach based on batch dynamic unfolded data is specially attractive. As commented, the addition of LMVs in a single model approach may be necessary to capture dynamics of a certain order and/or changing correlation structures. This makes difficult to interpret these models and to draw conclusions regarding the dynamic behaviour of the process. Moreover, these single phase models tend to be over-parameterized. In multi-phase models from batch dynamic unfolded data, dynamics are effectively modelled by adding a sufficient number of LMVs and changing correlation structures are modelled by using different sub-models for different periods of the batch. Therefore, contrarily to what happens with the single model approaches, the structure of the model corresponds to the dynamic nature of the process.

A careful study of the structure of the covariance matrices provides great insight on the ability of different approaches to model changing process dynamics. From this study, it is shown that the multi-phase strategy is the most flexible modelling approach for batch processes of those studied here. Nonetheless, its calibration may be a challenging task since the appropriate partition of the processes in several sub-models has to be identified. This can be done from expert knowledge, process analysis or automatic recognition.

Acknowledgements

Research in this area is partially supported by the Spanish government and the European Union (CICYT-FEDER DPI2005-01180 and CTM2005-06919-C03/TECNO) and by the FPU grants program, Secretaría de Estado de Educación y Universidades (Ministry of Education and Science, Spain), grant AP2003-0346.

The anonymous reviewers are acknowledged for their useful comments.

[§] Taking into account solely the part of the model applicable to new incoming data. For instance, in PCA the number of parameters is understood as the number of elements in the loadings matrix and in PLS the number of elements in the regressors matrix.

REFERENCES

1. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995; **3**7: 41–59.
2. Kassidas A, MacGregor JF, Taylor PA. Synchronization of batch trajectories using dynamic time warping. *AIChE J.* 1998; **4**4: 864–875.
3. Wold S, Kettanch N, Friden H, Holmberg A. Modelling and diagnostics of batch processes and analogous kinetic experiments. *Chemometrics Intell. Lab. Syst.* 1998; **4**4: 331–340.
4. Wise BM, Gallagher NB, Martin EB. Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. *J. Chemometrics* 2001; **1**5: 285–296.
5. Lu N, Gao F, Yang Y, Wang F. PCA-based modelling and on-line monitoring strategy for uneven-length batch processes. *Ind. Eng. Chem. Res.* 2004; **4**3: 3343–3352.
6. Fransson M, Folestad S. Real-time alignment of batch process data using COW for on-line process monitoring. *Chemometrics Intell. Lab. Syst.* 2006; **8**4: 56–61.
7. Tucker IR. The extension of factor analysis to three-dimensional matrices. In *Contributions to Mathematical Psychology*, Frederiksen N, Gulliksen H (eds). Holt, Rinehart and Winston: New York, 1964; 110–162.
8. Bro R. Multiway calibration. multi-linear PLS. *J. Chemometrics* 1996; **1**0: 47–61.
9. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **3**8: 149–171.
10. Smilde AK, Bro R, Geladi P. *Multi-Way Analysis, Application in the Chemical Sciences*. John Wiley & Sons: England, 2003.
11. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal components analysis. *AIChE J.* 1994; **4**0: 1361–1375.
12. Wold S, Geladi P, Esbensen K, Ohman J. Multi-way principal components and PLS analysis. *J. Chemometrics* 1987; **1**: 41–56.
13. Kourti T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemometrics* 2003; **1**7: 93–109.
14. Nelson PRC, Taylor PA, MacGregor JF. Missing data methods in PCA ans PLS: score calculations with incomplete observations. *Chemometrics Intell. Lab. Syst.* 1996; **3**5: 45–65.
15. Arteaga F, Ferrer A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *J. Chemometrics* 2002; **16**: 408–418.
16. Ündey C, Ertunç S, Çinar A. Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. *Ind. Eng. Chem. Res.* 2003; **4**2: 4645–4658.
17. Chen J, Liu K. On-line batch process monitoring using dynamic PCA and dynamic PLS models. *Chem. Eng. Sci.* 2002; **5**7: 63–75.
18. Camacho J, Picó J. Online monitoring of batch processes using multiphase principal component analysis. *J. Process Control* 2006; **1**0: 1021–1035.
19. Rännar S, MacGregor JF, Wold S. Adaptive batch monitoring using hierarchical PCA. *Chemometrics Intell. Lab. Syst.* 1998; **4**1: 73–81.
20. Ramaker H, Sprang ENM, Westerhuis JA, Smilde AK. Fault detection properties of global, local and time evolving models for batch process monitoring. *J. Process Control* 2005; **1**5: 799–805.
21. Lennox B, Montague GA, Hiden HG, Kornfeld G, Goulding PR. Process monitoring of an industrial fed-batch fermentation. *Biotechnol. Bioeng.* 2001; **7**4: 125–135.
22. Kosanovich KA, Dahl KS, Piovoso MJ. Improved process understanding using multiway principal component analysis. *Eng. Chem. Res.* 1996; **3**5: 138–146.
23. Ündey C, Çinar A. Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Syst. Mag.* 2002; **2**2: 40–52.
24. Lu N, Gao F, Wang F. Sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE J.* 2004; **5**0: 255–259.
25. Camacho J, Picó J. Multi-phase principal component analysis for batch processes modelling. *Chemometrics Intell. Lab. Syst.* 2006; **8**1: 127–136.
26. Aguado D, Ferrer A, Ferrer J, Seco A. Multivariate SPC of a sequencing batch reactor for wastewater treatment. *Chemometrics Intell. Lab. Syst.* 2007; **8**5: 82–93.

27. Dayal BS, MacGregor JF. Recursive exponentially weightes PLS and its applications to adaptive control and prediction. *J. Process Control* 1997; **7**: 169–179.
28. Qin SJ. Recursive PLS algorithms for adaptive data modeling. *Comput. Chem. Eng.* 1998; **2**2: 503–514.
29. Tipping ME, Bishop ChM. Mixtures of probabilistic principal component analysers. *Neural Comput., MIT Press*. 1999; **1**1: 443–482.
30. Zhao C, Wang F, Lu N, Jia M. Stage-based soft-transition multiple PCA modeling. *J. Process Control* 2007, DOI:10.1016/j.jprocont.2007.02.005
31. Camacho J, Picó J. Monitorización de procesos por lotes mediante PCA multifase. *Rev. Iberoam. Automat. Inf. Ind.* 2006; **3**: 78–91.
32. Camacho J, Picó J, Ferrer A. Multi-phase analysis framework for handling batch process data. Submitted to *J. Chemometrics* 2007.
33. Kourti T. Process analysis and abnormal situation detection: from theory to practice. *IEEE Control Syst. Mag.* 2002; **2**2: 10–25.
34. Lei F, Rotbøll M, Jørgensen SB. A biochemically structured model for Saccharomyces cerevisiae. *J. Biotechnol.* 2001; **8**8: 205–221.
35. Camacho J, Picó J, Ferrer A. Bilinear modelling of batch processes. Part II: PLS comparative. Submitted to *J. Chemometrics* 2007.

