Statistical analysis and online monitoring for handling multiphase batch processes with varying durations

Chunhui Zhao, Shengyong Mo, Furong Gao *, Ningyun Lu 1, Yuan Yao

Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

A R T I C L E   I N F O

Article history:
Received 15 September 2010
Received in revised form 10 February 2011
Accepted 15 April 2011
Available online 6 May 2011

Keywords:
Unequal multiphase batches
Uneven-length group
Subspace separation
Common and specific correlations
Variable-unfolding

A B S T R A C T

In the present work, statistical analysis and online monitoring is presented for handling uneven-length multiphase batch processes. Firstly, the irregular batches are classified into different uneven-length groups according to the changes of underlying characteristics. Then multi-source measurement data can be dealt with, each corresponding to one operation mode. The basic principle is that over different uneven-length groups, despite the uneven-length operation patterns, there are both similarity and dissimilarity to a certain extent among their underlying correlations. By an adequate decomposition, two different subspaces are separated, modeling the group-common and specific information respectively. Their corresponding confidence regions are constructed by searching similar patterns respectively. Accordingly, the online monitoring system is set up, which can track different types of variations closely. This analysis adds a detailed insight into the inherent nature of uneven-length multiphase batch processes. Its feasibility and performance are illustrated by a typical practical case with uneven cycles.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Batch processes are common in chemical, pharmaceutical, and food industries. Monitoring these batch processes is needed for various reasons such as safety, waste-stream reduction, consistency and quality improvement. Multivariate statistical techniques, such as multiway principal component analysis (MPCA) [1,2] and multiway partial least squares (MPLS) [3] were introduced by Nomikos and Macgregor for batch process modeling and monitoring. Since then, many applications and extensions of statistical batch process monitoring have been reported [4–12].

Generally, a vast amount of historical database on the measurement profiles is needed with completed batch runs that produced on-spec products. Subsequent to data acquisition, multivariate statistical analysis methods can thus be used to empirically model the successful historical operation batches. The variation within this data serves as reference distribution, against which the performance of independent new batches can then be compared. To apply the conventional statistical analysis methods, an implicit assumption is that the batch profiles are already time-scaled so that the operation events are synchronized over batches, resulting in equal durations. However, very often this assumption does not hold. The main reason for this wide range in batch time is related to the changes in operation conditions or control objectives. Take for instance the case when the end time depends on the amount of product produced or a quality constraint which has to be met. Commonly in these cases, data have to be properly aligned or equalized prior. Therefore, the data synchronization itself is a very crucial pre-processing step in the successful application of these multivariate methods.

Work has been carried out into various methods for equalizing batch lengths. In the simplest case, two batches have different lengths but the trajectories overlap in the common time part [13]. If this is the case, the problem can be very easily solved. Provided there are enough long batches, a model can be derived using the information from the long batches, while the absent part of trajectory of the shorter batches is treated as missing data [13]. The opposite is cutting the batches to a minimum length [14], and modeling focusing on this part, in which, however, the important information towards the end of longer batches would have been lost. Therefore, they are suitable only when the uneven-length problem is not serious and the main events have occurred in the common time part. Unfortunately, this is seldom satisfied since in the general case, the trajectories of the variables may have different shapes for most of the common time, showing different operation patterns. Regarding varying batch-to-batch process time, other handling methods [2,9,15] were presented including using rescaled batch time as a maturity index, tracking the batch progress with an indicator variable or using local batch time as the response vector in a PLS model. Nomikos and MacGregor [2] have suggested the use of another measured variable in place of
time index, termed an indicator variable. This variable progressed monotonically in time and had the same starting and ending value for each batch. Then the modeling and online application could be performed relative to its progress. Data resampling and interpolation are often needed to calculate the corresponding measurements at regular intervals of the selected indicator variable, which is done in a univariate way. This method could prove useful in chemical processes where process variables are recorded only when an event or change occurs rather than at fixed time intervals. However, since the aligned data are not the actual measured values of the process variables, the interpolations are not done carefully, it may not reflect the real correlations underlying process variables. Moreover, it is process specific and requires a process variable to be available that has similar properties to time. Such a variable may not exist in every chemical process that exhibits the appropriate mathematical properties, i.e., monotonic increasing. Two different warping techniques, dynamic time warping (DTW) and correlation optimization warping (COW), have previously been used as a method of pattern matching in speech recognition [16,17] and to correct chromatograms [18] for shifts in the time axis. They have received much attention in recent years to resolve the problem of uneven-length batch monitoring [19,20]. They can synchronize two trajectories by correcting a sample vector towards a reference one by appropriately translating, expanding, and contracting localized segments within the trajectories to achieve a minimum dissimilarity value between the two trajectories. Comparatively, DTW uses distance as a measure of similarity of two signals, while COW uses correlation coefficient as a similarity measure [21]. However, neither of them may uncover the test profile accurately enough by stretching or compressing it to the reference profile. 

In the present work, an enhanced statistical analysis and online monitoring strategy is developed for handling uneven-length multiphase batch processes. The objective is two-fold, referring to further process understanding and effective monitoring performance. The subject is those seriously uneven-length multiphase batches, which can be divided into different uneven-length groups. Instead of arbitrarily adopting global modeling or multiple modeling algorithms, their similarity and dissimilarity are first analyzed. Within the same group, the uneven problem is moderate over batches so that their underlying characteristics are similar. Over different uneven-length groups, the difference of underlying patterns may be more obvious resulting from more serious uneven-length problem. They are called specific correlations and might not be comprehensively described by a uniform model. On the other hand, it is also found that part of correlations will stay invariable and consistent over groups, which is called common ones here. The key is how to separate the two different types of correlations and model their associated variations respectively. Motivated by such cognition, a two-step basis vector extraction method [24], which was proposed to relate the inherent variable correlations over multiple data spaces, can be employed here as the basic modeling method. Following the development of a theoretical algorithm along with its property analysis in the previous work [24], this study addresses the potential of using the said algorithm to solve the uneven-length problem. For readability, a brief presentation of this algorithm is given in Appendix A. By this algorithm, within each uneven-length group, the original measurement space can be separated into two different subspaces. One is called the common subspace, revealing the similar underlying correlations over groups. The other is called the specific subspace enclosing the different correlations from one group to another. In comparison to the traditional statistical techniques, the separation of the common and specific correlations and the dual care paid to both are the major differences. They allow the batches to be uneven-length without the assumption that all batches show similar behaviors.

Fig. 1. Illustration of uneven-length problem for a two-phase batch process (over different batches \( \mathbf{X}_i \), batch index \( i = 1, 2, \ldots, l \), \( K_i \) is the length of the shortest batch process and \( K_L \) is the length of the longest batch process).

The rest of this paper is organized as follows. First, the uneven-length problem is stated for uneven-length multiphase batch processes. Subsequently, the proposed process analysis, statistical modeling and monitoring procedure are formulated in detail. The underlying principle support is revealed and its suitability and rationality are highlighted. In the following section, the application to a practical case is presented, in which, the presented recognition and argument are verified by the illustration results. Finally, conclusions are drawn in the last section.

2. Statement of uneven-length problem

In each batch run (batch index \( i = 1, 2, \ldots, l \)), assume that \( J \) process variables are measured online at \( k = 1,2, \ldots, K_i \) time instances throughout the operation cycle where the duration is not fixed in length, forming each irregular batch data set, denoted as \( \mathbf{X}_i(K_i \times j) \). Generally, there are two problems associated with modeling uneven-length batches. The first problem is simply the unequal length, meaning that irregular batches cannot be arranged as rows in a regular matrix. The second problem is that the local time or event of the batches is different, e.g., the batches may have reached a different maturity (such as conversion) at the same time.

For batch processes which typically consist of a sequence of steps, also known as sequential phases, the uneven-length problem will be more complex. Variability may occur in the durations of the phases so that events signifying the beginning or the end of a phase are generally misaligned in time over batches, as illustrated by batches \( \mathbf{X}_1 \) and \( \mathbf{X}_2 \) in Fig. 1. Even when batches externally have equal length, the process landmarks may be different, where, similar events happen at different time points with different phase durations, as indicated by batches \( \mathbf{X}_1 \) and \( \mathbf{X}_2 \) in Fig. 1. No matter whether the batch cycles are externally equal in duration or not, they may cover several uneven-length phases. In the general case, the trajectories of the variables have different shapes for most of the common time, as shown in Fig. 2 taking example for three uneven-length batches (A, B and C) against the average one. For better readability of the diagram, all the three illustrated batch profiles are plotted above the average one. The measurement point at the \( K \)th time in Batch A is more similar to the sample at the \( (K + 1) \)th time in Batch B and the sample at the \( (K + 1) \)th time in Batch C. It also tells us that different batches may arrive at the same underlying characteristics earlier or later. A shift of the data may be required when aligning or synchronizing the trajectories.

Clearly, to match the shape of the trajectories in these cases, the conventional batch-unfolding data arrangement is not proper since it simply places them in different rows and arranges them from the left in an arbitrary way. The calculated average trajectories over
batches do not stand for the true center resulting from the obvious alteration and shift of operation patterns over batches. Even after the phase data have been separated, modeling the batch-unfolding preprocessed batch data cannot reveal the normal batch-wise variability. Synchronization of operation patterns is thus necessary, in which, the process landmarks should be identified and then batch trajectories can be aligned with respect to landmark locations. Kaitha and Moore [22] presented a technique for the extraction of the time of occurrence of consistent feature in batch profiles. Once the event times were extracted, linear interpolation was used for scaling the time axis or to pad/chop the profiles. The approach, however, might be limited to a batch which is typically characterized by sharp features such as steps, ramps, peaks, etc., so that event times could be extracted clearly. Lu et al. [23] presented a phase-based sub-PCA modeling method for varying-length batches, where two PCA models were developed, one for the phase division and the other for process monitoring. They tried to focus on the batch-to-batch variation at each time, which, however, might not be proper when the uneven-length problem is serious. The variable trajectories may greatly mismatch against each other over batches at each time as shown in Fig. 2. The so-called batch-wise average trajectory may be meaningless any way as analyzed above even if it can be computed. Therefore, the modeling of each time-slice operation pattern fails to correctly reveal the normal batch-wise variations. Moreover, over irregular batches, the uneven-length cycles indicate that their underlying characteristics will be different more or less, which are actually a mixture of various operation patterns. It is natural to imagine that a global and representative model designed for all uneven-length batches may lose high resolutions for some ones. Especially for those seriously uneven-length batches, it is possible to find that the batch-wise variation might be even larger than the within-batch variation along time direction. A natural idea is thus to divide them into different uneven-length groups, in which, the underlying variable correlations remain similar within the same group but are more different over different groups. Moreover, despite of the varying cycles, the inherent nature of multiphase is still satisfied. Similar phase durations may be observed within the same group, as shown in Fig. 3. For example, in uneven-length Group u, the phase duration difference \( L_{u,\text{max}} - L_{u,\text{min}} \) is moderate. However, comparing two different groups \( u \) and \( u + 1 \), their phase duration difference \( L_{u+1,\text{max}} - L_{u,\text{min}} \) is more obvious and serious.

3. Methodology

In view of the modeling problems for uneven-length multiphase batch processes, the objective of the proposed method is two-fold, referring to further process understanding and effective monitoring performance. The flow diagram of the proposed method is illustrated in Fig. 4 where both model development as shown in Fig. 4(a) and online application as shown in Fig. 4(b) require careful treatment. In the modeling stage, based on the separation of different characteristics of different uneven-length batches, different models are designed and the monitoring system is set up with close confidence regions. During online application, the different uneven-length cases also should be distinguished so that the proper monitoring models can be adopted.

3.1. Phase and uneven-length group division

Based on the above analysis, before statistical modeling, both phase and uneven-length group information should be identified so that one representative model can be designed to enclose the similar operation behaviors of similar batches within the same phase. There are many phase partition methods [9,23–25–29] from different viewpoints and based on different principles. In the present work, assuming that no prior process knowledge is available, the phase information can be identified for each individual batch by clustering algorithm as what was done by Lu et al. [30] Certainly, the phase information can also be readily obtained if the practical process knowledge is known. Next, for uneven-length grouping, the dissimilarity of different uneven-length batches in the same phase will be checked. The uneven-length group division procedure is summarized as below:

First, for different batch data sets within the same phase, \( X_{i,p} \), \( (K_i \times f) \) (where subscripts \( i \) denotes batch and \( p \) denotes phase;
$K_{ip}$ is the number of samples of Batch $i$ in Phase $p$, they are stacked in variable-unfolding way. That is, all $X_{ip}(K_{ip} \times J) \ (i = 1, 2, . . . , I)$ are put together one after another, forming $X_p \left( \sum_{i=1}^{I} K_{ip} \times J \right)$ (where $I$ is the number of modeling batches). Then they are mean-centered and scaled to unit variance, designated as $\tilde{X}_p \left( \sum_{i=1}^{I} K_{ip} \times J \right)$. The effect of this preprocessing differs from that of batch-wise treatment. Direct mean centering of $X_p$ refers to subtracting grand phase means of variable trajectories over different irregular batches so that the time-varying trajectory variations within the same phase are made clear over batches. This type of unfolding and mean centering may be especially useful when a wide range of fluctuation trajectories are covered for some of the variables.

Then, after the data normalization, the different uneven batches are separated from each other and designated as $\tilde{X}_{ip}(K_{ip} \times J)$ ($i = 1, 2, . . . , I$). Fitted by PCA, $I$ phase-specific models $P_{ip}(J \times J)$ are obtained for different batches. They are then clustered. In this way, $U_p$ uneven-length groups are identified, each designated as $\tilde{X}_{up \cdot p} \left( \sum_{i=1}^{U_p} l_{up} K_{ip} \times J \right)$ (where $l_{up}$ is the number of batches belonging to the $u_p$th uneven-length group; subscript $u_p$ indicates the group index in Phase $p$, $u_p = 1, 2, . . . , U_p$). Without losing generality, although the total number of reference batches is sufficient, it may not be enough in each individual uneven-length group, which is not uncommon in real case. Generally, the grouping result can be directly associated with the varying phase durations. In each phase, the batches with moderate uneven problem will be collected into the same uneven-length group and those with serious uneven problem will be classified into different groups. It should be noted in different phases, the grouping results may be different, which is directly determined by the effects of irregular patterns on phase correlations.

### 3.2. Cross-group subspace separation

In each phase, from different uneven-length groups, multi-sources data sets are obtained. For subspace separation, the general

---

**Fig. 4.** Flow charts of the proposed uneven-length batch modeling and monitoring strategy (a) statistical model development and (b) online monitoring procedure.
idea is to analyze two different types of correlations from the cross-group viewpoint. The common operation patterns are the similar variable correlations over uneven-length groups. And the specific patterns are the correlations which are not shared by different groups. The multiset analysis algorithm [24], as shown in Appendix A, is used here to decompose the inherent variable correlations over different uneven-length groups and discriminate cross-group similarity and dissimilarity.

Within each phase over uneven-length groups, the common underlying phase basis vectors, \( \mathbf{P}^T_p(x \times R_p^p) \), are extracted (where \( R_p^p \) is the retained number of common basis in Phase \( p \), which is actually the same over different uneven-length groups; superscript \( c \) indicates the common information). Each data space of uneven-length group \((\bar{\bar{X}}_{up,p} \left( \sum_{i \in I_p} k_{i,p} x J \right))\) can thus be separated into two different parts, one common subspace \((\bar{\bar{X}}_{up,p} \left( \sum_{i \in I_p} k_{i,p} x J \right))\) and the residual part, called the specific subspace \((\bar{\bar{X}}_{up,p} \left( \sum_{i \in I_p} k_{i,p} x J \right))\), superscript \( s \) indicates the specific information:

\[
\bar{\bar{X}}_{up,p}^{c} = \bar{\bar{X}}_{up,p}^{c} + \bar{\bar{X}}_{up,p}^{s},
\]

where \( \bar{\bar{X}}_{up,p}^{c} = (P_c^T P_c p \times p) \) and \( \bar{\bar{X}}_{up,p}^{s} = (P_s^T P_s p \times p) \) are the linear combination coefficients corresponding to the common bases \((P_c^T)\). Actually, \( \mathbf{G}_{up,p} = (P_c^T P_c p \times p) \) is the orthogonal projector onto the column space of \( P_c^T \) and \( \mathbf{H}_{up,p} = I - \mathbf{G}_{up,p} = I - P_c^T P_c p \) is the anti-projector with respect to the column space of \( P_c^T \). Therefore, from another viewpoint, the two subspaces can also be obtained by projecting \( \bar{\bar{X}}_{up,p} \) onto the two different projectors respectively, \( \bar{\bar{X}}_{up,p}^{c} \) and \( \bar{\bar{X}}_{up,p}^{s} \). It is clear the two subspaces are orthogonal with each other since \( \bar{\bar{X}}_{up,p}^{c} (\bar{\bar{X}}_{up,p}^{s})^T = \bar{\bar{X}}_{up,p}^{c} \mathbf{G}_{up,p} (\bar{\bar{X}}_{up,p}^{s} \mathbf{H}_{up,p})^T = 0 \).

Moreover, from Eq. (1), the linear combination coefficients \( \mathbf{B}_{up,p}^{c} \left( \sum_{i \in I_p} k_{i,p} x R_p^p \right) \), where it should be noted that they are not guaranteed to be orthogonal with each other, are actually calculated by directly projecting \( \bar{\bar{X}}_{up,p} \) (or \( \bar{\bar{X}}_{up,p}^{c} \)) onto the common bases:

\[
\bar{\bar{X}}_{up,p}^{c} = \bar{\bar{X}}_{up,p}^{c} (P_c^T P_c p \times p) = \bar{\bar{X}}_{up,p}^{c} p = \mathbf{B}_{up,p}^{c}.
\]

In the common subspace, the objects enclose the same underlying correlations by the linear combinations of one subset of common bases. These common bases actually denote the most initial and basic correlations underlying this subspace, which can be used as the uniform statistical model. While in the left subspace, the variable correlations cover no information associated with the common bases and thus are more specific and unique to each uneven-length group, revealing their dissimilarity. Therefore, different model structures should be designed over different specific subspace. Here, the traditional PCA decomposition is performed to separate the systematic and residual information in each specific subspace:

\[
\bar{\bar{X}}_{up,p}^{s} = \bar{\bar{X}}_{up,p}^{s} (I - P_c^T p \times p) = \bar{\bar{X}}_{up,p}^{s} (I - P_c^T p \times p) P_p^T p \times p ,
\]

\[
\bar{\bar{X}}_{up,p}^{s} = \bar{\bar{X}}_{up,p}^{s} P_p^T p \times p ,
\]

\[
\bar{\bar{X}}_{up,p}^{s} = \bar{\bar{X}}_{up,p}^{s} (I - P_c^T p \times p).
\]

where \( \bar{\bar{X}}_{up,p}^{s}(J \times R_p^p) \), the specific PCA loadings, reveal the major specific variation directions and \( R_p^p \) is the retained principal component (PC) number. \( \bar{\bar{X}}_{up,p}^{s} \left( \sum_{i \in I_p} k_{i,p} x J \right) \) are the specific residuals, also the final modeling errors.

Generally, \( \bar{\bar{X}}_{up,p} \left( J \times R_p^p \right) \), should be more different over uneven-length groups if those common bases are all excluded. From another viewpoint, resulting from the basic PCA model, those PCA loadings can also be expressed as one form of linear combination of the original measurement, \( \bar{\bar{X}}_{up,p} = \bar{\bar{X}}_{up,p}^{s} (T_{up,p} (T_{up,p}^T T_{up,p})^{-1} \). They are thus can be regarded as one type of basis vectors. It is clear that the specific bases \((\bar{\bar{X}}_{up,p}^{s})\) are orthogonal to those common bases \((\bar{\bar{X}}_{up,p}^{c})\) based on Eq. (1). Therefore, the group-specific scores \( T_{up,p} \left( \sum_{i \in I_p} k_{i,p} x R_p^p \right) \) can also be obtained by directly projecting the original group measurement \( \bar{\bar{X}}_{up,p} \) onto \( \bar{\bar{X}}_{up,p}^{c} p \times p\).

In summary, the underlying characteristics of each uneven-length group are formulated in two subspaces as below:

\[
\bar{\bar{X}}_{up,p} = \bar{\bar{X}}_{up,p}^{c} + \bar{\bar{X}}_{up,p}^{s} = \mathbf{B}_{up,p}^{c} (P_p^T p \times p) = T_{up,p} (T_{up,p}^T T_{up,p})^{-1} + E_{up,p}^{c}.
\]

\[
\bar{\bar{X}}_{up,p} = \bar{\bar{X}}_{up,p}^{c} + \bar{\bar{X}}_{up,p}^{s} = \mathbf{B}_{up,p}^{c} (P_p^T p \times p) + T_{up,p} (T_{up,p}^T T_{up,p})^{-1} + E_{up,p}^{c} = \bar{\bar{X}}_{up,p}^{c} \mathbf{G}_{up,p}^{c} + \bar{\bar{X}}_{up,p}^{s} + E_{up,p}^{c}.
\]

(1)

(4)

where for each uneven-length group in Phase \( p \), the complete systematic model, \( \mathbf{O}_{up,p} = [P_p^T p \times p] \). , describes both common and specific systematic correlations. Correspondingly, \( \mathbf{B}_{up,p}^{c} \) and \( \mathbf{T}_{up,p} \) are the associated variations along \( \mathbf{P}_c^T \) and \( \mathbf{P}_s^T \), respectively.

Then the SPE values of the final residuals after the dual explanation of both common and specific systematic models can be calculated readily as below:

\[
\mathbf{SPE}_{\ell,up,p} = \mathbf{e}_{\ell,up,p}^T \mathbf{e}_{\ell,up,p} = \sum_{i \in I_p} k_{i,p} x J ,
\]

(5)

where \( \mathbf{e}_{\ell,up,p} \) is the row vector of \( \sum_{i \in I_p} k_{i,p} x J \), and subscript \( \ell = 1, 2, \ldots, \sum_{i \in I_p} k_{i,p} x J \).

3.3. Construction of local confidence region

Based on the calculation of systematic scores and residuals, three different monitoring statistics can be designed. The key is how to construct their confidence regions. For all the irregular reference batches, in each phase \( p \), the corresponding scores in the common and specific subspaces are collected as \( \mathbf{B}_{up,p}^{c}, \mathbf{B}_{up,p}^{s} \), and \( \mathbf{B}_{up,p}^{c}, \) and \( \mathbf{B}_{up,p}^{s} \) respectively. The extracted underlying systematic information \( \mathbf{B}_{up,p}^{c}, \mathbf{B}_{up,p}^{s} \) actually captures the major evolution trend of every operation cycle along time direction in each group as shown in Fig. 5(a). They might not be proper to be directly used for process monitoring because the operation trajectories along the time direction fluctuate over a wide range and the global confidence region may be too loose. To get close confidence regions of systematic variations, the normal batch-wise variations should be focused on and those scores which reveal the similar event information should be collected. Here, for both common and specific scores, the confidence regions are locally constructed using a query-based approach by searching those similar scores in reference database, in which the basic analysis unit is time-slice.
Moreover, the time-slice covariances are also calculated
\[\Sigma_{c_{up,p}} = \{\Sigma_{c_{1,up,p}}^c, \Sigma_{c_{2,up,p}}^c, \ldots, \Sigma_{c_{k_{up,p},up,p}}^c, \ldots, \Sigma_{c_{K_{up,p},up,p}}^c\}\]
\[\Sigma_{s_{up,p}} = \{\Sigma_{s_{1,up,p}}^s, \Sigma_{s_{2,up,p}}^s, \ldots, \Sigma_{s_{k_{up,p},up,p}}^s, \ldots, \Sigma_{s_{K_{up,p},up,p}}^s\}\]

At certain sampling interval where only one score vector is available, unit matrix is used as its covariance. They are then used by the designed search procedure for confidence region construction of common scores and specific scores. The detailed procedure is shown in Appendix B. Moreover, the confidence limit for SPE values can be readily constructed as indicated by the selection result of specific scores since the residuals go through the dual explanation by the same common and specific models as those specific features. The residuals are deemed to follow normal variations and thus SPE can be described by a weighted Chi-squared distribution [2].

By the search procedure described in Appendix B, at each sampling time in each uneven-length group, the similar operation patterns are collected and more close control regions can thus be defined for three different monitoring statistics. It is a different treatment from previous work.

3.4. Online monitoring strategy

For the new observation vector \(x_{new}(1 \times 1)\), it is first normalized by the grand mean and variance of the current phase. The corresponding common systematic variation scores are calculated by projecting \(x_{new}\) onto the unified common model, \(b_{new}^c = P_T x_{new}\). It may belong to any one uneven-length group \((up)\). To identify its accurate affiliation, calculate the Mahalanobis distance between the new common score and each group, which, actually, is the calculation of \(T^2\) statistic:
\[T^2_{up,new} = (b_{new}^c - b_{k,up,p}^c)\sum_{k_{up,p}}^{-1}(b_{new}^c - b_{k,up,p}^c),\]

where \(b_{k,up,p}^c\) is the average of the common scores enclosed in the constructed local confidence region at the same time \(k\) in each uneven-length group and \(\Sigma_{k,up,p}^c\) is the corresponding covariance structure.

Based on the indication of Mahalanobis distance, select the uneven-length group where the distance value stays well below the confidence limit as the affiliation group \((u_p)\). Then the specific score for the new observation can be calculated by adopting the phase model in this group, \(t_{new} = P_{T,u_p} x_{new}\) and its Hotelling-\(T^2\) can be calculated based on the associated local confidence region:
\[T^2_{new} = (t_{new} - t_{k,up,p}^c)\sum_{k_{up,p}}^{-1}(t_{new} - t_{k,up,p}^c),\]

where \(t_{k,up,p}^c\) is the average of the specific scores within the prepared confidence region at the same time \(k\) in the chosen \(u_p\) th uneven-length group and \(\Sigma_{k,up,p}^s\) is the corresponding covariance structure.

The final residual and its SPE value can then be calculated as below:
\[e_{new} = x_{new} - t_{new} = (I - \Omega_{up,p} \Omega_{up,p}^T)x_{new}\]
\[SPE_{new} = e_{new}^T e_{new}\]

Moreover, for online monitoring, one key point is about how to judge the current phase so that the proper phase model can be adopted. As mentioned before, within each uneven-length group, although the uneven-length problem is moderate, the durations may still be irregular over different batches. It means process time is insufficient for judging which phase model should be adopted. Here assume that the shortest and longest lengths of the irregular
phases in the same uneven-length group \( (4) \) are \( L_{\text{up}}^{P, \text{min}} \) and \( L_{\text{up}}^{P, \text{max}} \) respectively as shown in Fig. 3. Then the online application can be implemented according to the following two cases:

(a) For the data within the range \([1, L_{\text{up}}^{P, \text{min}}]\), it is clear that the data belong to the current phase definitely. So the current phase model can be directly adopted to calculate the monitoring statistics.

(b) For the data belonging to the range \([L_{\text{up}}^{P, \text{min}} + 1, L_{\text{up}}^{P, \text{max}}]\), the current sample may still lie in the current phase or it has entered the next phase. It is important to distinguish the normal phase switch and process fault. For the normal data within the current phase, the monitor statistics calculated by the current phase models can be well enclosed by the confidence regions. For the normal data entering the next phase, the current phase models fail to explain its underlying variable correlations and will give out-of-control monitoring statistics. The data, however, can be well fitted by the next phase models, which means by adopting the next phase models, the false alarms will be eliminated. For the process fault sample, neither of the current and the next phase models can describe it well and the out-of-control indication cannot be removed. Therefore, the phase switch will be readily distinguished from abnormal behaviors.

4. Experiment results

Injection molding [26,37], a key process in polymer processing, transforms polymer materials into various shapes and types of products. A typical injection molding process consists of three operation phases, injection (or filling) of molten plastic into the mold, packing-holding of the material under pressure, and cooling of the plastic in the mold until the part becomes sufficiently rigid for ejection. Besides, plastication takes place in the barrel in the early cooling phase, where polymer is melted and conveyed to the barrel front by screw rotation, preparing for next cycle.

Experiments are designed based on the real instrumented reciprocating-screw injection-molding machine settled in our laboratory. It can be set as a typical uneven-length multiphase batch process, in which, the filling phase duration is not fixed but rather depends on the injection velocity. Obviously, a lower injection velocity results in more filling time and thus a longer batch with more process data. Although it is possible for the real injection-molding machine to repeat each cycle rigidly with a high degree of automation, this operation is intentionally implemented as an uneven-length batch process to demonstrate the proposed method. Here, the injection velocity is artificially set to change from 16 to 40 mm/s, involving four typical velocity values: 16, 24, 32 and 40 mm/s. Moreover, around 24 mm/s, the injection velocity is set to be 22–26 with varying sample points 96–84, simulating the moderate uneven problem which can be classified into the same uneven-length group as analyzed later. The maximal filling phase duration is 128 sample intervals corresponding to injection velocity 16 mm/s and the minimal filling duration is 58 associated with 40 mm/s. Therefore the difference of filling duration is 70 samples, about 1.2 times of the minimal filling length. This is different from the case used in Lu’s method [23], where the analysis subject only allowed the injection velocity to change from 22 to 26 mm/s, limited to moderate uneven-length problem. In our present work, clearly, the uneven problem is more serious, which thus will result in different underlying characteristics to a certain extent. Here for simplicity, except the filling phase, the other phases are controlled to have exactly the same duration, that is, they have equal trajectories.

The material used in this work is high-density polyethylene (HDPE). Nine process variables are selected for modeling as shown in Table 1, which can be collected online with a set of sensors. The packing-holding time is fixed at 3 s; the cooling time is set at 15 s. Resulting from the varying filling phase, the irregular batch cycles have 663–733 sampling intervals. Forty normal batch runs are used for model building, where ten cycles are collected for each kind of injection velocity (16, 22–26, 32, 40 mm/s).

First, focusing on each batch respectively, the process duration is preparatorily partitioned into five main phases, in which, operation

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nozzle temperature</td>
<td>℃</td>
</tr>
<tr>
<td>2</td>
<td>Nozzle pressure</td>
<td>Bar</td>
</tr>
<tr>
<td>3</td>
<td>Screw stroke</td>
<td>mm</td>
</tr>
<tr>
<td>4</td>
<td>Screw velocity</td>
<td>mm/s</td>
</tr>
<tr>
<td>5</td>
<td>Injection pressure</td>
<td>Bar</td>
</tr>
<tr>
<td>6</td>
<td>Plastication pressure</td>
<td>Var</td>
</tr>
<tr>
<td>7</td>
<td>SV1 opening</td>
<td>%</td>
</tr>
<tr>
<td>8</td>
<td>SV2 opening</td>
<td>%</td>
</tr>
<tr>
<td>9</td>
<td>Cavity pressure</td>
<td>Bar</td>
</tr>
</tbody>
</table>

Table 1 Description of the used process variables.

Fig. 6. Illustration of variable correlations over uneven-length groups in filling phase (a) the first two common bases; (b) the first specific basis.
time information is included so that process samples are consecutive within the same phase. Generally, the phase division result is deemed to be consistent with the real four physical operation phases: filling, packing-holding, plastication, and cooling. Moreover, one short phase occurs between the packing-holding and plastication phases. Then the phase model is developed for each batch and over batches the phase models are clustered so that different uneven-length groups are classified. Here, in filling phase, they are partitioned into four uneven-length groups with about 128, 90, 70, 58 sampling intervals respectively. For other equal phases, analyzing their phase PCA models, it is found that they have the similar correlations and can be archived into the same uneven-length group. It means that generally different injection velocity values do not influence the underlying correlations in later phases. For the filling phase, the extracted first two common bases using the algorithm shown in Appendix A are shown in Fig. 6(a), revealing the similar variable correlations over uneven-length groups. The basis coefficients corresponding to the fourth process variable (injection velocity) is small, revealing that the injection velocity has little contribution to the common variable correlations. This well agrees with the real situation that the uneven-length filling duration results from different injection velocity settings. Moreover, under the influence of different injection velocity values, the SV1 and SV2 openings (Variables 7 and 8) also have little contributions. After the extraction of the common model, different specific models are developed, focusing on revealing the differences over groups. The specific bases are shown in Fig. 6(b) taking the first one for instance. The basis coefficient corresponding to Injection Velocity is larger compared with those shown in Fig. 6(a), reflecting the real situation. The two-dimensional specific score \((t_1 - t_2)\) planes are shown in Fig. 7 for all four uneven-length groups in the filling phase, where they are more different from each other.

For construction of local confidence region, with respect to common feature, the similar ones can be selected over groups and time. Taking example for the first twenty sampling intervals in filling phase within the third uneven-length group where the injection velocity is 32 mm/s, the selection result is shown in Table 2. For example, for the 20th sampling time-slice in Group 3, the common pattern is more similar to the one at later time in Group 1 because Group 1 has larger phase duration. In the current study, for simplicity, the confidence limits for the selected regions are all approximately defined from the normal distribution. That is, \(T^2\) by

![Fig. 7. The specific scores plane \((t_1 - t_2)\) in filling phase over different uneven-length groups.](image)

### Table 2
Common pattern selection result in the filling phase in the second uneven-length group.

<table>
<thead>
<tr>
<th>Process time in Group 3</th>
<th>First selected pattern</th>
<th>Second selected pattern</th>
<th>Third selected pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group index</td>
<td>Time index</td>
<td>Group index</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>37</td>
<td>3</td>
</tr>
</tbody>
</table>
the group affiliation is definitely fixed and the operation status is also checked in terms of systematic information in both subspaces. With time evolving, the affiliation judgment result will be more stable and accurate. Finally, the residual information is checked and shown in Fig. 8(c), which also indicates that the current batch is operating normally in filling phase.

Within the second uneven-length group, in the filling phase, the shortest phase length is 84 intervals corresponding to injection velocity 26 mm/s and the longest phase length is 96 intervals resulting from 22 mm/s. Then the monitoring is implemented according to two cases: before the shortest phase length [1,84] and within the period [85,96]. Before the shortest filling phase length [84], the online monitoring is implemented by adopting the filling phase model. For the data within the period [85,96], to differentiate a process fault from a normal phase switch, those data that exhibit out-of-control monitoring statistics using the filling phase models are hypothetically regarded as the beginning of the packing-holding phase. They are then checked by recalculating the scores by adopting the next phase models. The online monitoring results are shown in Fig. 9(a) using the filling phase models. Generally, all the monitoring statistics are well below the control limits before 90th sampling intervals, indicating that the batch is operating within the filling phase and free of any process abnormality. After that the filling phase models fail to explain the current operation patterns and give higher statistics beyond the control limits for all monitoring statistics. Then one may want to differentiate the
cause of the alarms, a real process fault or the normal phase switch. The data will be renormalized accordingly and the second phase model is then adopted to check the data within region [91,96], as shown in Fig. 9(b). Since the operation durations are equal for all batches in packing-holding phase, subPCA model [23] is enough to model their underlying characteristics. Clearly, it yields in-control statistics, which means the current batch has entered the second phase after 90th sampling time. This well agrees with the real case that the filling velocity of the current batch is 24 mm/s, resulting in a 90-sample filling phase.

Moreover, one process fault is introduced by withdrawing the injection pressure which will directly affect the operation patterns in filling phase. The monitoring result is shown in Fig. 10, revealing good fault detection performance. Generally, clear and stable alarms are revealed especially by the $T^2$ monitoring system based on specific scores, which means the abnormal behavior mainly influences the specific systematic information. This is subject to the real case that the fault is rendered by changing injection pressure (Variable 5), which is the major source of specific variable correlations as shown in Fig. 6(b).

5. Conclusions

In the present work, an alternative method is provided for analysis of uneven-length multiphase batches. The strengths of the proposed strategy lie in not only the effective monitoring but also the appealing analysis results and comprehension for uneven-length problem. By uneven-length batch grouping and subspace separation, the underlying correlations of uneven-length batches can be analyzed more comprehensively. Higher-resolution models and closer confidence regions are thus set up by collecting similar patterns. It precludes the problem of correlation distortion resulting from trajectory modification in conventional algorithms. Illustration to injection molding process illustrates its effectiveness.

Acknowledgements

This work is supported in part by the China National 973 program (2009CB320603).

Appendix A.

In the proposed modeling method, two key points are the design of different statistical models for the better characterization of different variable correlations and the construction of close confidence regions for the improved monitoring performance. Different uneven-length groups are separated in each phase, preparing mul-
multiple data sets. The multiset data analysis algorithm is shown in this appendix, which is used to discriminate cross-group similarity and dissimilarity of the inherent variable correlations. A detailed search procedure is designed as shown in Appendix B, which is used to collect the similar operation patterns at each sampling time in each uneven-length group for the construction of close local confidence regions.

In case of multiset measurement data, \( X(N_i \times J) \), the set-to-set interrelation may refer to the common structures in variable correlations. In each measurement space, it is always possible to find out a subset of samples, which are representative enough to the other samples and can substitute all samples by their linear combinations. The major underlying variable correlations in the original measurement space are also represented by them and the associated distribution variances, revealing their respective significances. They are called subbasis vectors here, which are employed to evaluate the similarity and dissimilarity over multiple sets.

Since any subbasis in each dataset space, \( p_j(j = 1, 2, \ldots, J) \), must lie in the span of the input observations, there exist linear combination coefficients \( a_{ij} = [a_{ij}, a_{ij}^2, \ldots, a_{ij}^N] \), such that

\[
p_j = \sum_{n=1}^{N_i} a_{ij}^n x_n^i = X^T a_j^i
\]  

(A1)

That is, each subbasis vector \( p_j \) is actually a linear function of the original observations in each dataset.

The degree of similarity of subbasis vectors should be measured in terms of “how close with each other over sets”. However, it would be complicated if all set-to-set interrelationships are simultaneously and directly evaluated. Here, the computation trick by Carroll [31] is borrowed, in which, a third-party canonical variate was introduced in their GCA algorithm. Then in our method, the simultaneous similarity assessment of variable correlations over sets can be achieved through the introduction of a global and common basis vector, \( p_g \). It can be regarded as the supplementary and pseudo \((C+1)\)st subbasis vector and should approximate all \( C \) subbases as close as possible. That is, these real subbasis vectors, which are correlated with each other as close as possible, or speaking more exactly, as common as possible over sets, can be comprehensively described and even substituted by the global basis.

To figure out the common bases, a two-step extraction procedure is designed. In the first step, the common subbases are preparatorily computed from the original measurement data and then in the second step they can be further condensed and refined by enhancing their correlations. Different optimization objectives and constraints are used in the two steps, which both come down to the simple analytic solutions of constrained optimization problems.

A.1. The first-step basis extraction

During the first-step basis extraction, we define it in terms of finding a \( J \)-dimensional global basis \( \{p_g\} \) together with different linear combinations of the observations comprising each of \( C \) collective data sets with the cost function and certain constraints as below:

\[
\max R^2 = \max \sum_{i=1}^{C} (p_i^T X^T a_i)^2
\]  

s.t. \( p_g^T p_g = 1 \) \( a_i^T a_i = 1 \)  

(A2)

Using a Lagrange operator, the optimization problem finally leads to a simple analytic solution:

\[
\sum_{i=1}^{C} (X^T X)p_g = \lambda_g p_g
\]  

(A3)

\[
Q p_g = \lambda_g p_g
\]

This is a standard algebra problem. At the request of the maximal objective function value, i.e., the largest \( \lambda_g \), analytically, the solution leads to the eigenvalue decomposition on the sum of subset covariances, \( Q = \sum_{i=1}^{C} (X^T X) \).

The subbasis vector is calculated by:

\[
p_j = X^T a_j = \sqrt{\frac{1}{\lambda_i}} X^T X p_g
\]  

(A4)

where the parameter \( \lambda_i \) can be calculated by \( p_g^T X^T X p_g = \lambda_i \).

Ordered, \( R \) number of global basis vectors can be derived by Eq. (A3) in accord with the descending \( \lambda_g \), resulting in the same number of subbasis vectors calculated using Eq. (A4) in each dataset. This decomposition summarizes and compresses the underlying cross-set common variable correlations into a new subspace spanned by \( R \) subbases within each dataset, \( \{p_j \} \).

A.2. The second-step basis extraction

To get the cross-set common subbases which are really close correlated, correlation analysis index should be used instead of the covariance index. In comparison with the optimization function and constraints shown in Eq. (A2), the second-step basis extraction is designed by constructing and solving a different optimization problem. It is implemented on the basis of the first-step analysis result \( \{p_j \} \) and the aim is to maximize the mean square correlations:

\[
\max R^2 = \max \sum_{i=1}^{C} (\bar{p}_i^T \bar{p}_g)^2 = \max \sum_{i=1}^{C} (\bar{a}_i^T \bar{a}_g)^2
\]  

s.t. \( \bar{p}_i^T \bar{p}_g = 1 \) \( \bar{a}_i^T \bar{a}_g = 1 \)  

(A5)

Using a Lagrange operator, its solution also comes down to a standard algebra problem:

\[
\sum_{i=1}^{C} (\bar{p}_i^T \bar{p}_g)^2 \bar{p}_i = \lambda_g \bar{p}_g
\]  

(A6)

\[
S \bar{p}_g = \lambda_g \bar{p}_g
\]

Therefore, the optimization problem finally leads to a simple analytical solution, i.e., \( p_g \) should be the eigenvector of \( S = \sum_{i=1}^{C} (\bar{p}_i^T \bar{p}_g)^2 \) corresponding to the largest eigenvalue \( \lambda_g \).
The subbasis vector is calculated by:
\[
p_i = \hat{p}_g^T a_i = \frac{1}{\sqrt{\lambda_i}} \hat{p}_g^T (\hat{P} \hat{P}^T)^{-1} \hat{p}_g
\]
(A7)
where the sub-optimal objective parameter \(\lambda_i\) can be calculated by
\[
p_g^T \hat{P}^T (\hat{P} \hat{P}^T)^{-1} \hat{P} \hat{p}_g = \lambda_i.
\]
In turn, \(R\) number of global basis vectors are retained and can construct a global basis subspace \(P_g(R \times J)\). Correspondingly, \(C\) sub-basis subspaces, \(P_g(R \times J)\), are also derived, which are actually the projected ones from \(P_g(R \times J)\) onto \(\hat{P}^T\).

Appendix B.

B.1. Search procedure for close confidence region

B.1.1. Input

(i) The current query points at time \(k\) of phase \(p\) in group \(u_p\), i.e., the time-slice common and specific features, \(B_{k,up,p} = (\hat{b}_k^T, \hat{c}_k^T, \ldots, \hat{b}_k^T, \hat{c}_k^T, \ldots, \hat{b}_{up,p}^T, \hat{c}_{up,p}^T)\) and their centers as well as covariances:
\[
\hat{b}_{k,up,p}^T, \hat{b}_{k,up,p}^T, \ldots, \hat{b}_{up,p}^T, \hat{c}_{up,p}^T, \hat{c}_{up,p}^T\text{ and }\Sigma_{k,up,p}^2, \Sigma_{up,p}^2.
\]
(ii) The proposed reference score time-slices in different subspaces over groups \((u_p)\) within the same phase \((p)\):
\[
B_{k,up,p}^T = (\hat{b}_{1,up,p}^T, \hat{b}_{2,up,p}^T, \ldots, \hat{b}_{K,up,p}^T, \ldots, \hat{b}_{K,up,p}^T, \hat{c}_{up,p}^T, \ldots, \hat{c}_{up,p}^T, \hat{c}_{up,p}^T, \ldots, \hat{c}_{up,p}^T, \hat{c}_{up,p}^T).
\]
(iii) The number counter \(L\) as its initial value. The required vector number to enclose the query point, \(L_q\), where it is set to satisfy two or three times of the number of the original process variables to ensure a reliable statistical model as commonly practiced [32].

B.1.2. Output

The similar patterns around each query point as well as their confidence limits: \(B_{k,up,p}^T\) and \(T_{k,up,p}^T\); \(\text{Ctr}_{k,up,p}^T\) and \(\text{Ctr}_{k,up,p}^T\).

B.2. Step (1), for the systematic common information

(a) For the query time-slice scores \(B_{k,up,p}^T\) in one uneven-length group, to choose the similar common score time-slice, calculate the Mahalanobis distance between each candidate time-slice and the query one: \(d_{ISK} = 1/R_{k,up,p} \sum_{i} I_{k,up,p}(b_i - \hat{b}_{k,up,p}^T, \Sigma_{k,up,p}^{-1}(b_i - \hat{b}_{k,up,p}^T)\),

where subscript \(i\) denotes the available batches \(R_{k,up,p}\) in each candidate common score time-slice. The average is used to generally evaluate their similarity.

(b) The time-slice scores \(B_{k,up,p}^T\) with the least distance value are selected and added to the query ones. \(L = L + 1\).

(c) Do iteration until the selected number of patterns in this local region \(L\) is no less than \(L_q\). Then by combining the selected patterns and the original query ones, calculate their new center and covariance, and the Hotelling-\(T^2\) values can be readily computed. The confidence limit is then designed by certain distribution [2,33] or assigning a value that is slightly larger than the maximum reference one.

(d) From the phase beginning in each uneven-length group, implement the above procedure one by one, the similar common patterns can be collected at each time in any group for confidence region construction.

B.3. Step (2), for the specific systematic information

The selection procedure is similar to that for the common systematic information. It is worth remarking that the search region is limited to the neighboring time regions in the current group because different systematic models are used over groups, revealing their different underlying correlations. It is different from what has been done in Step (1), where the possible search region may span over groups because a unified phase model has been used for the calculation of common scores.

B.3.1. Discussion and remarks

(1) In comparison with the conventional calculation method, here, the analysis subject of the selection algorithm is the reduced feature subspace instead of the original process measurement space. The underlying variable correlations of samples have been taken into consideration potentially and evaluated by those predefined models so that those samples whose correlations are best fit for the query point can be chosen. Moreover, it uses Mahalanobis distance index to evaluate the similarity between the current time-slice and the possible candidate one, which is actually the calculation way Hotelling-\(T^2\) takes. In this way, the \(T^2\) chart can detect small shifts and deviations from normal region.

(2) As shown in the search steps, the number of finally chosen samples is \(L = \sum I_{k,up,p}\) where \(I_{k,up,p}\) is the number of available batches in those chosen time-slices. Therefore, the more the batches within each uneven-length group, the narrower the time spanning of the search region, which means the collected similar patterns can be deemed to well follow batch-wise normal distribution. On the contrary, the less the batches, the wider the search region stretches along time direction, which means the variation represented by the selected patterns will be larger since time-varying dynamics are included. Therefore, the number of batches within each uneven-length group will determine the construction accuracy of the confidence region. This can be comprehended from the following two extreme cases:

(a) The number of batches within each uneven-length group is enough. The confidence region can be well defined by each individual time-slice within the same group, which well follows \(F\)-distribution for Hotelling-\(T^2\) statistics [2,33].

(b) The number of batches within each uneven-length group is extremely small, which means that the normal batch-wise variations at each time within the same group cannot be comprehensively represented. The \(F\)-distribution is compromised when collecting the similar patterns over groups or time.

References