Enhanced Process Comprehension and Quality Analysis Based on Subspace Separation for Multiphase Batch Processes

Chunhui Zhao and Furong Gao
Dept. of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong SAR

Dapeng Niu and Fuli Wang
College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, P. R. China

DOI 10.1002/aic.12275
Published online May 4, 2010 in Wiley Online Library (wileyonlinelibrary.com).

Phase-based subpartial least squares (subPLS) modeling algorithm has been used for online quality prediction in multiphase batches. It strictly assumes that the ***X–Y*** *correlations are identical within the same phase so that they can be defined by a uniform regression model. However, the accuracy of this precondition has not been theoretically checked when put into practical application. Actually it does not always agree well with the real case and may have to be rejected for some practical processes. In the present work, it corrects the "absolute similarity" of subPLS modeling by a more general recognition that only one part of the underlying correlations are time-wise common within the same phase while the other part are time-specific, which is referred to as "partial similarity" here. Correspondingly, a two-step phase division strategy is developed, which separates the original phase measurement space into two different parts, the common subspace and uncommon subspace. It is only in the common subspace where the underlying* ***X–Y*** *correlations are similar, a phase-unified regression model can be extracted for online quality prediction. Moreover, based on the subspace separation, offline quality analyses are conducted in both subspaces to explore their respective cumulative manner and contribution in quality prediction. The strength and efficiency of the proposed algorithm are verified on a typical multiphase batch process, injection molding.* © 2010 American Institute of Chemical Engineers *AIChE J,* 57: 388–403, 2011

Keywords: multiphase batch processes, partial similarity, subspace separation, common patterns, phase representability, cumulative manner

Introduction

Data-based multivariate calibration methods have been widely used to establish a quantitative relationship between process measurement (**X**) and quality property (**Y**). Accurate qualitative and quantitative calibration analysis may help avoiding cumbersome and costly chemical measurements. In practice, calibration modeling and analysis can often be accomplished with familiar, conventional statistical techniques,[1–6] such as multiple linear regression (MLR), principal component regression (PCR), canonical correlation analysis

Correspondence concerning this article should be addressed to F. Gao at kefgao@ust.hk.

(CCA), and partial least squares (PLS). Among them, latent variable (LV)-based methods play a dominating role. They often work well because variable collinearity is typically strong and partly redundant over a large number of measured variables. Thus, modeling the variable correlation pattern allows shrinking of the original data space into a lower-dimensional feature subspace. Fewer uncorrelated LVs can be defined to comprehensively represent the original input variables and used to build a quantitative regression relationship with the concerned quality properties.

The subject of calibration modeling and quality interpretation arouses new issues and demand specific solutions when it refers to multiphase (MP) batch processes, where various phases generally operate orderly under the domination of different physical phenomena, revealing different effects on the final qualities. MPLS model[7] uses process variables over the entire batch course as the input, which reveals well the time correlations throughout the cycle and thus shows efficiency for batches which progress in a more cumulative manner. However, for MP batches, it is generally deemed that if the data are handled in a single matrix, the effect of one segment tends to be hidden more or less by the influence of another. The resulting tribulation is that the hidden effect could be useful in quality-concerned process analysis and control. It is commonly accepted that in MP cases, more underlying information can be explored by dividing the data into meaningful blocks either by the types of variables or by the part of the process they originate from and building multiple specific models instead of single modeling of all data. The effect of each block can be seen and thus more comprehensive process understanding can be expected. Considering that the phase multiplicity is an inherent nature of many batch processes, various strategies[8–25] have been reported and can be put into practical process monitoring and quality prediction. Camacho and Pico proposed a MP algorithm[8,9] for automatic phase identification so that each segment of the batch can be well approximated by a linear PCA model with acceptable nonexplained variance. Later, they[10] extended the MP algorithm from PCA to PLS and lagged variables were included to model the variable dynamics for online quality prediction. Liu et al.[11] applied PLS to monitor the interphase relation of a two-stage batch process, where in postanalysis of abnormalities, it could clarify whether root causes were from previous phase operation or due to the changes of interphase correlations. Moreover, in previous work, the phase-based modeling techniques have been extended to solve different practical problems in process monitoring,[12–14] such as problems of uneven length and limited reference batches. Zhao et al.[15] presented a two-level MP calibration modeling strategy to probe the phase-wise local and cumulative effects on quality interpretation and prediction. Yao and Gao[16] gave an overview of MP statistical analysis methods for process analysis, monitoring, quality prediction, and online quality improvement, where different types of phase divisions and modeling strategies were analyzed and discussed.

From the viewpoints of online and offline quality analyses, respectively, different application purposes determine their differences in both phase identification and model design. For phase-based online quality analysis, subPLS[18,19] algorithm is the major research direction, which allows one to unveil the time-varying quality variation information with no data complement. For phase-based offline quality analysis, two major parallel lines of thought can be employed. One[15,21] is to model each phase separately by MPLS algorithm. In the present work, it is called phase-based MPLS (P-MPLS) in comparison with the original MPLS.[7] And the other is to model the variable correlation within each phase under the influence of other phases by multiblock PLS (MBPLS).[22–26] The objective is to extract the covarying systematic dynamics between phases for quality prediction which can not been explored when each phase is analyzed individually.

The subPLS modeling algorithm was first developed by Lu and Gao[18] for online quality prediction at each sampling time. It was based on such a presupposition that despite the time-varying batch operation trajectory, the correlations between process and quality variables should remain similar within the same phase. Therefore, indicated by the changes of underlying characteristics, a batch cycle could be divided into appropriate number of phases with little prior process knowledge. Each phase covered a series of similar and time-consecutive patterns which could be modeled by a simple phase-representative PLS model. For online application, at each sampling time, according to its affiliated phase, the corresponding phase-representative model was adopted and a realtime quality prediction was obtained with no data estimation. Critical-to-quality phases were also identified in which the online quality prediction could be accepted with confidence. Moreover, a phase-based online quality control scheme[19] was reported by the same authors based on the real-time quality prediction information. The control action was taken realtime to steer the process inputs and compensate for the quality loss in the past when the end-product quality was deviating from the desired value as indicated by online prediction. Moreover, it should be pointed out that the PLS modeling algorithm based on phase-specific variable unfolding[20] is essentially equivalent to subPLS as it impliedly accepts the same precondition as subPLS. Here for simplicity, it is also archived into the subPLS modeling framework.

P-MPLS and MBPLS methods can be used as the major offline quality analysis strategies to apprehend the phase behaviors and their effects on quality prediction. They have been recommended in cases where the number of variables is large and can be separated into conceptually meaningful blocks, which can help to localize the regression relationships in a decentralized manner. However, P-MPLS only explored the variable collinearity isolated in each phase but overlooked the relationships between different phases. Compared with P-MPLS, the advantage of MBPLS was mainly to allow for easier interpretation of both the roles of each smaller meaningful block and the integrated contribution of all blocks. Qin et al.[25] have made a comprehensive review of multiblock algorithms and reported a successful application to an industrial polyester film process. Duchesne and MacGregor[27] developed a pathway MBPLS algorithm which was based on the consideration that the process variables affecting quality only over a specific period of time were fairly difficult to identify with the final quality measurements alone. It incorporated the quality measurements collected during the course of a batch, which were regarded to be helpful to isolate the local effects of variable trajectory changes. However, for practical batch processes, it is often found that the intermediate quality measurements can not be readily obtained.

Comparing the above analyzed phase-based regression modeling strategies, subPLS algorithm focuses on the individual effect of each sampling time on quality, thus accounting for neither the within-phase nor between-phase correlations. Especially, it was based on a strict and ideal assumption that the correlations between process and quality were quite similar within the same phase, which was a prelude to further analysis. However, the correctness of such a precondition has never been examined prior to modeling. From the practical viewpoint, the ideal ‘‘absolute similarity’’ may not be well satisfied, which thus directly influences the accuracy of subPLS model and online prediction performance. P-MPLS and MBPLS modeling ideas do not care whether the underlying characteristics within the same phase are identical or not, in which, a general phase regression relationship can be always derived. However, the identification of multiple blocks/phase has not been automatically accomplished and mainly depended on the prior process knowledge. Moreover, they are less effective in probing into the realtime quality variation information compared with subPLS modeling strategy.

In this study, an enhanced process understanding and quality analysis strategy is developed for MP batches. The underlying operation patterns within each phase are deeply investigated according to their correlations with the product quality. It relaxes the strict assumption of subPLS by a more general recognition that within the same phase the underlying **X–Y** correlations not only share similarity to a certain extent (called common patterns here) but also show dissimilarity over time (called uncommon patterns here). For accuracy, it is called ‘‘partial similarity’’ here in contrast with the ‘‘absolute similarity’’ in the original subPLS algorithm. From a mathematical viewpoint, a statistical evaluation index is designed to examine the phase-specific similarity and to what extent it is satisfied, absolutely or partially. Accordingly, a further phase identification approach is developed which can simultaneously divide the batch cycle into different proper phases and separate two subspaces in each phase: one is called the common subspace with similar underlying correlations and regression relationships; and the other is termed the uncommon subspace with different quality-relevant characteristics. Then a more representative phase model can be designed in the common subspace for online application by excluding the uncommon patterns. Besides, the difference of the common patterns and uncommon ones in the cumulative manner and effects is also distinguished. It is analyzed by P-MPLS and MBPLS performed in the two different subspaces respectively, which gives improved model interpretation and additional underlying phase information. Here it should be noted that as what has been clearly stated in previous work,[16] the divided blocks/segments, which more exactly, may be called ‘‘modeling phases,’’ are defined from the viewpoint of statistical meaning. They focus on reflecting the changes of the underlying correlation characteristics[17] and may be different from the real physical ‘‘operating phases.’’ For simplicity, they are uniformly called ‘‘phase’’ in the present work unless otherwise noted.

Methodology

In each batch run, assume that J process variables are measured online at $k = 1, 2, \ldots, K$ time instances throughout the batch and J_y quality variables are obtained offline. Then, process observations collected from similar I batches can be organized as a three-way array $\underline{\mathbf{X}}$ $(I \times J \times K)$ and a corresponding quality matrix $\mathbf{Y}(I \times J_y)$ as shown in Figure 1a. At each time, the means of each column are subtracted to approximately eliminate the main nonlinearity. Each variable is scaled to unit variance to handle different measurement units, thus giving each equal weight. In the present work, the batches are of equal length without special declaration so that the specific process time can be used as an indicator.

Selection of regression algorithm

PLS algorithm has been widely used to approximate the regression relationship between **X** and **Y**. However, its objective is to maximize their covariance, which thus also models their respective variations. Large covariance may not necessarily mean strong correlation. It is possible that a pair of principal directions in the two spaces have high covariance merely because the associated distribution variances are large. When the **X** space contains large amount of quality-uninformative process variations, resulting from their contamination, PLS often requires many LVs to achieve good fitting, which leads to more complex model structure and difficulty in model interpretation.

To improve PLS regression model, various preprocessing methods have been reported. Among them, variable selection[28,29] can directly reduce the model dimension by removing those quality-irrelevant input variables. In contrast, orthogonal signal correction (OSC),[30–32] as a feature extraction technique, tries to remove quality-irrelevant underlying components. Alternatively, CCA[3,33,34] is well-suited for relating two data tables. Unlike PLS, it directly exploits and maximizes the correlation instead of the covariance and thus inherently ignores the **Y**-irrelated variations in **X**. However, as the measurement variables are often high-dimensional and closely correlated, directly applying CCA to the raw input space will lead to an ill-conditioned problem because it involves the calculation of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. That is why CCA is not so popular as PLS in practical application. Yu and MacGregor[35] developed a PLS-CCA algorithm, in which, as a postprocessing, CCA was implemented on PLS LVs to further condense them. In this way, it avoided the rank-deficiency problem and got rid of the pseudo quality-relevant variations so that a parsimonious regression model was obtained with the same prediction ability as the standard PLS model. Their work has showed its advantages over OSC-based preprocessing approach. In the present work, the PLS-CCA algorithm will be adopted as the basis of calibration analysis.

Phase representability evaluation

As stated in Introduction, the key point of our algorithm is that it adopts the concept of ‘‘partial similarity’’ within the same phase, which is quite different from the ‘‘absolute similarity’’ of the original subPLS. Therefore, it is necessary to check whether how the ‘‘partial similarity’’ precondition is satisfied or not by the given batch process. Here, the different characteristics of common and uncommon patterns should be first analyzed so that an evaluation criterion can

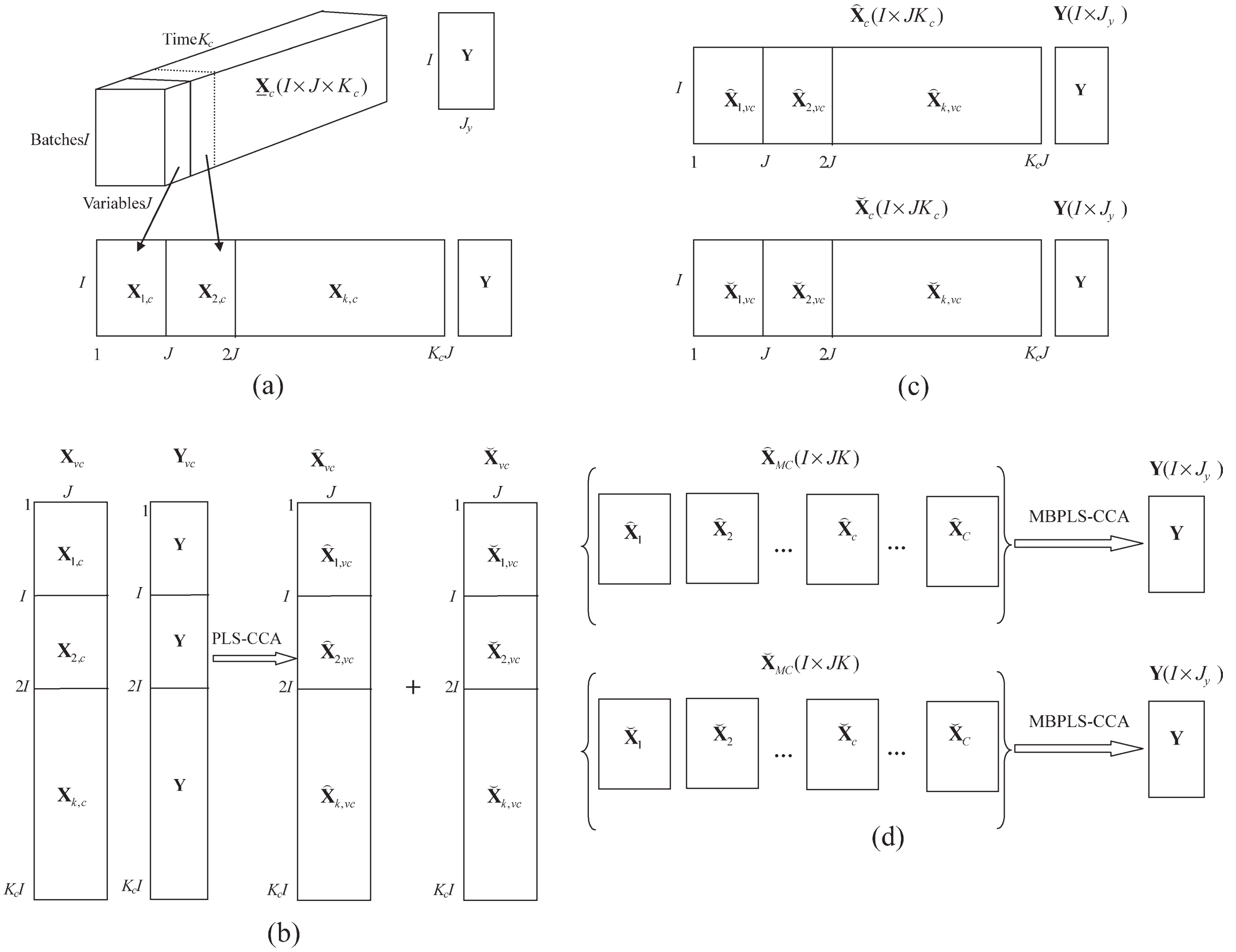

TimeK_c
Batches$I$
Variables$J$
$\underline{\mathbf{X}}_c(I\times J\times K_c)$
I
$\mathbf{Y}$
J_y
$\mathbf{X}_{1,c}$
$\mathbf{X}_{2,c}$
$\mathbf{X}_{k,c}$
1
J
$2J$
K_cJ
(a)
$\mathbf{X}_{vc}$
$\mathbf{Y}_{vc}$
$\overset{\frown}{\mathbf{X}}_{vc}$
$\breve{\mathbf{X}}_{vc}$
I
$2I$
K_cI
PLS-CCA
+
$\overset{\frown}{\mathbf{X}}_{1,vc}$
$\overset{\frown}{\mathbf{X}}_{2,vc}$
$\overset{\frown}{\mathbf{X}}_{k,vc}$
$\breve{\mathbf{X}}_{1,vc}$
$\breve{\mathbf{X}}_{2,vc}$
$\breve{\mathbf{X}}_{k,vc}$
(b)
$\overset{\frown}{\mathbf{X}}_c(I\times JK_c)$
$\mathbf{Y}(I\times J_y)$
$\breve{\mathbf{X}}_c(I\times JK_c)$
(c)
$\overset{\frown}{\mathbf{X}}_{MC}(I\times JK)$
$\breve{\mathbf{X}}_{MC}(I\times JK)$
$\overset{\frown}{\mathbf{X}}_1$
$\overset{\frown}{\mathbf{X}}_2$
$\overset{\frown}{\mathbf{X}}_c$
$\overset{\frown}{\mathbf{X}}_C$
$\breve{\mathbf{X}}_1$
$\breve{\mathbf{X}}_2$
$\breve{\mathbf{X}}_c$
$\breve{\mathbf{X}}_C$
...
MBPLS-CCA
(d)

Figure 1. Illustration of the phase-based regression modeling scheme in *c*th phase.

(a) Batch-wise unfolding and data normalization, (b) subspace separation and modeling for online quality prediction, (c) within-phase cumulative analysis, and (d) between-phase cumulative analysis.

be defined for guidance. In detail, it can be mathematically formulated as follows.

In each phase, performing regression analysis focusing on variable-unfolding data set $\mathbf{X}_{\mathrm{vc}}$ and $\mathbf{Y}_{\mathrm{vc}}$ (as shown in Figure 1b) by PLS-CCA algorithm, a A_{vc}-dimensional phase-based PLS-CCA weights matrix, $\mathbf{R}_{\mathrm{vc}}$, can be always obtained, from which, one will check whether the common part ($\overset{\frown}{\mathbf{R}}_{\mathrm{vc}}$) and uncommon part ($\breve{\mathbf{R}}_{\mathrm{vc}}$) are mixed together. The corresponding PLS-CCA LVs ($\mathbf{T}_{\mathrm{vc}}$), which may mix both the common and uncommon ones ($\overset{\frown}{\mathbf{T}}_{\mathrm{vc}}$ and $\breve{\mathbf{T}}_{\mathrm{vc}}$), can be calculated as below:

$$\begin{aligned}\mathbf{T}_{\mathrm{vc}} &= \mathbf{X}_{\mathrm{vc}}\mathbf{R}_{\mathrm{vc}}\\ \overset{\frown}{\mathbf{T}}_{\mathrm{vc}} &= \mathbf{X}_{\mathrm{vc}}\overset{\frown}{\mathbf{R}}_{\mathrm{vc}}\\ \breve{\mathbf{T}}_{\mathrm{vc}} &= \mathbf{X}_{\mathrm{vc}}\breve{\mathbf{R}}_{\mathrm{vc}}\end{aligned} \tag{1}$$

where, the phase-based PLS-CCA weights model, $\overset{\frown}{\mathbf{R}}_{\mathrm{vc}}(J \times \overset{\frown}{A}_{\mathrm{vc}})$ is intrinsically doubly controlled by PLS and CCA weights, $\overset{\frown}{\mathbf{R}}{}^{\mathrm{pls}}_{\mathrm{vc}}(J \times \overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}})$ and $\overset{\frown}{\mathbf{R}}{}^{\mathrm{cca}}_{\mathrm{vc}}(\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}} \times \overset{\frown}{A}_{\mathrm{vc}})$, in which, $\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}}$ is the number of PLS weights directions and $\overset{\frown}{A}_{\mathrm{vc}}$ denotes the CCA dimension, which is also the dimension of the final PLS-CCA model.

Only in the common subspace, the regression relationships stay similar within the same phase and change over phases. Therefore, for the within-phase common patterns, a consistent phase model ($\mathbf{R}_{\mathrm{vc}}$) can well approximate them, i.e., $\overset{\frown}{\mathbf{R}}_{\mathrm{vc}}$ has higher phase representability and will be well similar to those obtained by directly regression modeling focusing on $\mathbf{X}_k$ and $\mathbf{Y}$. In contrast, the underlying **X–Y** correlations in the uncommon subspace are time-varying even in the same phase, which, thus, can not be comprehensively described by any phase-representative regression model. That is, $\mathbf{R}_{\mathrm{vc}}$ is far from reflecting the real time-varying regression weights relationship and will be quite different from those obtained by directly regression modeling focusing on $\mathbf{X}_k$ and $\mathbf{Y}$. More exactly, it should be called pseudo model. Based on the above analysis, for the common operation patterns, because of their within-phase similarity, the common time-slice LVs, $\mathbf{T}_{k,\mathrm{vc}}$, which are split from $\mathbf{T}_{\mathrm{vc}}$ or actually calculated by $\mathbf{T}_{k,\mathrm{vc}} = \mathbf{X}_k\mathbf{R}_{\mathrm{vc}}$ as shown in Eq. 1, will satisfy the orthogonality well, so that their covariance will be near to diagonal or as diagonal as possible:

$$\widehat{\mathbf{R}}_{\mathrm{vc}}^{\mathrm{T}}\frac{\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k}{I-1}\widehat{\mathbf{R}}_{\mathrm{vc}}=\frac{\widehat{\mathbf{T}}_{k,\mathrm{vc}}^{\mathrm{T}}\widehat{\mathbf{T}}_{k,\mathrm{vc}}}{I-1}=\widehat{\Lambda}_{k,\mathrm{vc}} \tag{2}$$

where, $\widehat{\Lambda}_{k,\mathrm{vc}}$ is a $\widehat{A}_{\mathrm{vc}}$-dimensional diagonal matrix, whose elements actually denote the distribution variance information along $\widehat{A}_{\mathrm{vc}}$ common weight directions at each time.

In contrast, for the uncommon operation patterns, resulting from their within-phase dissimilarity, the uncommon time-slice LVs, $\breve{\mathbf{T}}_{k,\mathrm{vc}}$, which are split from $\breve{\mathbf{T}}_{k,\mathrm{vc}}$ or actually calculated by $\breve{\mathbf{T}}_{k,\mathrm{vc}}=\mathbf{X}_k\breve{\mathbf{R}}_{k,\mathrm{vc}}$ as shown in Eq. 1, thus do not satisfy the orthogonality. That is, their covariance can not be transformed to a diagonal form by a uniform phase model.

To check whether the ''partial similarity'' within the same phase can be satisfied, it actually can be evaluated by checking whether uncommon patterns exist in the time-slice LVs. Based on the above analyses, resulting from the different characteristics between common and uncommon patterns within the same phase, the uncommon ones can be distinguished by assessing the deviation from diagonality of those time-slice LV covariances. The evaluation index is defined as below:

$$\varphi_k(\mathbf{T}_{k,\mathrm{vc}}^{\mathrm{T}}\mathbf{T}_{k,\mathrm{vc}})=\left|\mathrm{diag}(\mathbf{T}_{k,\mathrm{vc}}^{\mathrm{T}}\mathbf{T}_{k,\mathrm{vc}})\right|\bigg/\left|\mathbf{T}_{k,\mathrm{vc}}^{\mathrm{T}}\mathbf{T}_{k,\mathrm{vc}}\right| \tag{3}$$

where, || denotes the determinant operation and diag(**F**) is the diagonal matrix having the same diagonal elements as **F**.

The fact that φ_k is a reasonable measure of deviation from diagonality for a positive definite symmetric matrix can be seen from Hadamard's inequality[36]:

$$\left|\mathbf{T}_{k,\mathrm{vc}}^{\mathrm{T}}\mathbf{T}_{k,\mathrm{vc}}\right|\le\left|\mathrm{diag}(\mathbf{T}_{k,\mathrm{vc}}^{\mathrm{T}}\mathbf{T}_{k,\mathrm{vc}})\right| \tag{4}$$

where, equality exists only if **F** is diagonal. Therefore, φ_k (**F**) ≥ 1 holds, with equality exactly when **F** is diagonal. Actually, φ_k (**F**) increases monotonically as **F** is continuously ''inflated'' from diag(**F**) to **F**. This has been demonstrated by Flury and Gautschi in their work.[37]

Then the average of all time-slice φ_k values within the same phase can be used to comprehensively evaluate the phase representability of regression model:

$$\varphi_{\mathrm{c}}=\frac{1}{K_{\mathrm{c}}}\sum_{k\in c}\varphi_k \tag{5}$$

where, K_{c} is the current phase duration.

If ''absolute similarity'' is satisfied, the within-phase time-slice LVs ($\mathbf{T}_{k,\mathrm{vc}}$) are all common patterns. That is, after proper phase division, the phase model is representative enough so that the within-phase time-slice LVs well satisfy the diagonality of covariances ($\varphi_k \approx 1$), resulting in φ_{c} approximating 1. When only ''partial similarity'' is accepted, i.e., the within-phase time-slice LVs ($\mathbf{T}_{k,\mathrm{vc}}$) cover both common and uncommon patterns, failing to satisfy the diagonality of time-slice covariances (φ_k (**F**) $\gg 1$) and resulting in φ_{c} greatly larger than 1. However, the original phase space can be separated into two parts, the phase-specific common subspace and the uncommon subspace. Only in the common subspace which encloses the similar patterns, φ_{c} approximates 1; whereas in the uncommon subspace, φ_{c} is greatly larger than 1. When no within-phase similarity is approved, no common subspace can be figured out to achieve $\varphi_{\mathrm{c}} \approx 1$. Based on the above analysis, how the current process satisfies the within-phase similarity, partially or absolutely, can be readily determined by checking whether a common subspace can be separated as well as the value of the statistical index defined in Eq. 5.

Phase-based subspace separation

Based on the definition of evaluation index for phase representability, a complete phase division algorithm is designed here, which can simultaneously achieve the phase division and subspace separation. Its goal is to find the modeling phases in the batch process which can be better approximated by models with sufficient representability. It is sequentially executed in two steps: a preliminary clustering step and a further evaluation and subdivision step which is shown in Appendix A. In the first step, subPLS clustering algorithm[18] is implemented, which can find an initial solution (a partition of the batch duration into C groups) and use it as the preliminary basis for further analysis. Then in the second step, the statistical index defined in Eqs. 8 and 10 will be used to evaluate the initial clusters and correct them if they are far from describing the phase nature. It can tell one how the first-step clustering result is, how representative the resulting regression models are and which clusters require further subdivision. Especially, it can automatically judge whether the common subspace really exists.

Based on the two-step division result, online quality prediction model can be derived in the common subspace within each phase. Combined with Eq. 1, the common phase loadings ($\widehat{\mathbf{P}}_{\mathrm{vc}}$) are calculated and the common and uncommon subspaces ($\widehat{\mathbf{X}}_{\mathrm{vc}}$ and $\breve{\mathbf{X}}_{\mathrm{vc}}$) can be separated as below:

$$\begin{aligned}
\widehat{\mathbf{P}}_{\mathrm{vc}}^{\mathrm{T}}&=\left(\widehat{\mathbf{T}}_{\mathrm{vc}}^{\mathrm{T}}\widehat{\mathbf{T}}_{\mathrm{vc}}\right)^{-1}\widehat{\mathbf{T}}_{\mathrm{vc}}^{\mathrm{T}}\mathbf{X}_{\mathrm{vc}}=\Lambda_{\mathrm{vc}}^{-1}\widehat{\mathbf{R}}_{\mathrm{vc}}^{\mathrm{T}}\mathbf{X}_{\mathrm{vc}}^{\mathrm{T}}\mathbf{X}_{\mathrm{vc}}\\
\widehat{\mathbf{X}}_{\mathrm{vc}}&=\widehat{\mathbf{T}}_{\mathrm{vc}}\widehat{\mathbf{P}}_{\mathrm{vc}}^{\mathrm{T}}\\
\breve{\mathbf{X}}_{\mathrm{vc}}&=\mathbf{X}_{\mathrm{vc}}-\widehat{\mathbf{X}}_{\mathrm{vc}}=\mathbf{X}_{\mathrm{vc}}\left(\mathbf{I}-\widehat{\mathbf{R}}_{\mathrm{vc}}\widehat{\mathbf{P}}_{\mathrm{vc}}^{\mathrm{T}}\right)
\end{aligned} \tag{6}$$

where, Λ_{vc} is a diagonal matrix with equal element $(I-1)K_{\mathrm{c}}$. Here it should be pointed out the phase LVs $\widehat{\mathbf{T}}_{\mathrm{vc}}$ have a unity variance resulting from the corresponding requirement of canonical variables in CCA. However, for each time-slice common LVs ($\widehat{\mathbf{T}}_{k,\mathrm{vc}}$) which are split from $\widehat{\mathbf{T}}_{\mathrm{vc}}$, they are not guaranteed to have unit variance. It means although the underlying correlations in the common subspace within the same phase are similar but the associated variance levels may be distinct over time.

Regression modeling for online quality prediction

In the separated common subspace, a simple J-dimensional phase-based regression model can be readily derived by performing PLS-CCA on data set $\left\{\widehat{\mathbf{X}}_{\mathrm{vc}},\mathbf{Y}_{\mathrm{vc}}\right\}$ in each phase:

$$\begin{aligned}
&\overset{\frown}{\mathbf{Q}}{}^{\mathrm{T}}_{\mathrm{vc}} = \left(\overset{\frown}{\mathbf{T}}{}^{\mathrm{T}}_{\mathrm{vc}}\overset{\frown}{\mathbf{T}}_{\mathrm{vc}}\right)^{-1}\overset{\frown}{\mathbf{T}}{}^{\mathrm{T}}_{\mathrm{vc}}\mathbf{Y}_{\mathrm{vc}} = \Lambda_{\mathrm{vc}}^{-1}\overset{\frown}{\mathbf{R}}{}^{\mathrm{T}}_{\mathrm{vc}}\mathbf{X}^{\mathrm{T}}_{\mathrm{vc}}\mathbf{Y}_{\mathrm{vc}}\\
&\overset{\frown}{\Theta}_{\mathrm{vc}} = \overset{\frown}{\mathbf{R}}_{\mathrm{vc}}\overset{\frown}{\mathbf{Q}}{}^{\mathrm{T}}_{\mathrm{vc}}
\end{aligned} \tag{7}$$

where, $\overset{\frown}{\mathbf{Q}}_{\mathrm{vc}}$ is the phase loadings for qualities and $\overset{\frown}{\Theta}_{\mathrm{vc}}$ is the regression coefficients matrix.

Therefore, at each time within the current phase, a real-time quality prediction can be obtained online:

$$\hat{\mathbf{Y}}_{k,\mathrm{vc}} = \mathbf{X}_k\overset{\frown}{\Theta}_{\mathrm{vc}} = \mathbf{X}_k\overset{\frown}{\mathbf{R}}_{\mathrm{vc}}\overset{\frown}{\mathbf{Q}}{}^{\mathrm{T}}_{\mathrm{vc}} = \overset{\frown}{\mathbf{T}}_{k,\mathrm{vc}}\overset{\frown}{\mathbf{Q}}{}^{\mathrm{T}}_{\mathrm{vc}} \tag{8}$$

The online quality predictions will be time-varying within the same phase, which may be caused by the measurement noises, modeling errors and particularly the different variance levels of $\mathbf{T}_{k,\mathrm{vc}}$.

Only in critical-to-quality phases, the online quality forecast results are credible. Here, the identification of critical phases uses the same method presented in our previous work,[18] i.e., checking the prediction accuracy of the phase models. Moreover, based on the obtained realtime quality information in critical phases, predicted deviation in the earlier period may be compensated in its following period by taking proper quality control action online.

Despite of the realtime strength, online quality analysis, however, only employs the common patterns and overlooks the process variation information in the uncommon subspace. Although the uncommon patterns are not suitable to online quality analysis, they actually have important contribution to qualities. Moreover, it fails in revealing the variable correlations along time directions. Some questions naturally arise: What will happen when the time-wise correlations are taken into consideration? Whether the common patterns still have the similar effects on qualities when they are put together within each phase? And how do the uncommon patterns cumulatively act on qualities? Moreover, under the influences of the within- and between-phase correlations respectively, will the cumulative phase behaviors play differently? And what will their differences are? On the basis of subspace separation, all the above mentioned will be answered by comparatively analyzing their cumulative effects from both within- and between-phase viewpoints.

Within-phase cumulative analysis

To reveal the cumulative effects of each individual phase, phase-based multiway regression algorithm is used here to conduct quality analyses in the common and uncommon subspaces respectively so that their respective effects can be revealed. Then their joint contribution will be explored by combining their quality prediction results at the end of each phase.

By Eq. 6, we have separated the original measurement space into two subspaces, $\overset{\frown}{\mathbf{X}}_{\mathrm{vc}}$ and $\breve{\mathbf{X}}_{\mathrm{vc}}$. They are unfolded batchwise and then the data units for within-phase cumulative analysis can be obtained in both subspaces as shown in Figure 1c:

$$\begin{aligned}
&\mathbf{X}_{\mathrm{c}}(I \times JK_{\mathrm{c}}) = \overset{\frown}{\mathbf{X}}_{\mathrm{c}} + \breve{\mathbf{X}}_{\mathrm{c}}\\
&\overset{\frown}{\mathbf{X}}_{\mathrm{c}}(I \times JK_{\mathrm{c}}) = \left[\overset{\frown}{\mathbf{X}}_{1,\mathrm{vc}}, \overset{\frown}{\mathbf{X}}_{2,\mathrm{vc}}, \ldots, \overset{\frown}{\mathbf{X}}_{K_{\mathrm{c}},\mathrm{vc}}\right]\\
&\breve{\mathbf{X}}_{\mathrm{c}}(I \times JK_{\mathrm{c}}) = \left[\breve{\mathbf{X}}_{1,\mathrm{vc}}, \breve{\mathbf{X}}_{2,\mathrm{vc}}, \ldots, \breve{\mathbf{X}}_{K_{\mathrm{c}},\mathrm{vc}}\right]
\end{aligned} \tag{9}$$

Within-phase cumulative analysis regression models are then designed in both subspaces:

In the common subspace:

$$\begin{aligned}
&\overset{\frown}{\mathbf{T}}_{\mathrm{c}} = \overset{\frown}{\mathbf{X}}_{\mathrm{c}}\overset{\frown}{\mathbf{R}}_{\mathrm{c}}\\
&\hat{\overset{\frown}{\mathbf{X}}}_{\mathrm{c}} = \overset{\frown}{\mathbf{T}}_{\mathrm{c}}\overset{\frown}{\mathbf{P}}{}^{\mathrm{T}}_{\mathrm{c}}\\
&\overset{\frown}{\mathbf{E}}_{\mathrm{c}} = \overset{\frown}{\mathbf{X}}_{\mathrm{c}} - \hat{\overset{\frown}{\mathbf{X}}}_{\mathrm{c}} = \overset{\frown}{\mathbf{X}}_{\mathrm{c}}\left(\mathbf{I} - \overset{\frown}{\mathbf{R}}_{\mathrm{c}}\overset{\frown}{\mathbf{P}}{}^{\mathrm{T}}_{\mathrm{c}}\right)\\
&\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}} = \overset{\frown}{\mathbf{T}}_{\mathrm{c}}\overset{\frown}{\mathbf{Q}}{}^{\mathrm{T}}_{\mathrm{c}}
\end{aligned} \tag{10}$$

where, $\overset{\frown}{\mathbf{R}}_{\mathrm{c}}$ is weights matrix; $\overset{\frown}{\mathbf{P}}_{\mathrm{c}}$ is loadings matrix for $\overset{\frown}{\mathbf{X}}_{\mathrm{c}}$; $\overset{\frown}{\mathbf{Q}}_{\mathrm{c}}$ is loadings matrix for qualities. $\breve{\mathbf{T}}_{\mathrm{c}}$ are the common phase scores; $\hat{\overset{\frown}{\mathbf{X}}}_{\mathrm{c}}$ are the modeled process variations and $\breve{\mathbf{E}}_{\mathrm{c}}$ are the residuals; and $\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}$ are the quality analysis results.

In the uncommon subspace:

$$\begin{aligned}
&\breve{\mathbf{T}}_{\mathrm{c}} = \breve{\mathbf{X}}_{\mathrm{c}}\breve{\mathbf{R}}_{\mathrm{c}}\\
&\hat{\breve{\mathbf{X}}}_{\mathrm{c}} = \breve{\mathbf{T}}_{\mathrm{c}}\breve{\mathbf{P}}{}^{\mathrm{T}}_{\mathrm{c}}\\
&\breve{\mathbf{E}}_{\mathrm{c}} = \breve{\mathbf{X}}_{\mathrm{c}} - \hat{\breve{\mathbf{X}}}_{\mathrm{c}} = \breve{\mathbf{X}}_{\mathrm{c}}\left(\mathbf{I} - \breve{\mathbf{R}}_{\mathrm{c}}\breve{\mathbf{P}}{}^{\mathrm{T}}_{\mathrm{c}}\right)\\
&\hat{\breve{\mathbf{Y}}}_{\mathrm{c}} = \breve{\mathbf{T}}_{\mathrm{c}}\breve{\mathbf{Q}}{}^{\mathrm{T}}_{\mathrm{c}}
\end{aligned} \tag{11}$$

where, the regression model and statistics are nominated one by one in a consistent way with those defined in Eq. 10 but in a different subspace.

The complete within-phase cumulative analysis results can then be obtained by combining the predictions from both subspaces:

$$\begin{aligned}
&\left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}, \hat{\breve{\mathbf{Y}}}_{\mathrm{c}}\right] \xrightarrow{\mathrm{PLS-CCA}(\mathbf{W}_{\mathrm{yc}})} \mathbf{Y}\\
&\hat{\mathbf{Y}}_{\mathrm{c}} = \left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}, \hat{\breve{\mathbf{Y}}}_{\mathrm{c}}\right]\mathbf{W}_{\mathrm{yc}} = \hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}\overset{\frown}{\mathbf{W}}_{\mathrm{yc}} + \hat{\breve{\mathbf{Y}}}_{\mathrm{c}}\breve{\mathbf{W}}_{\mathrm{yc}}
\end{aligned} \tag{12}$$

where, $\mathbf{W}_{\mathrm{yc}}(4 \times 2) = \begin{bmatrix}\overset{\frown}{\mathbf{W}}_{\mathrm{yc}}\\ \breve{\mathbf{W}}_{\mathrm{yc}}\end{bmatrix}$ is the weights derived by regression analysis between the quality predictions in both subspaces $\left(\left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}, \hat{\breve{\mathbf{Y}}}_{\mathrm{c}}\right]\right)$ and the final quality measurements ($\mathbf{Y}$ $(I \times J_y)$.

Between-phase cumulative analysis

The previous phase-based multiway modeling only explores the within-phase cumulative effects. For MP batch processes, besides the time correlations within each isolated phase, different phases also covary with each other. Conventional MPLS uses the measurements throughout the entire process duration as the input unit and thus can reveal the joint role of multiple phases. However, it loses its eye in the local phase information. As analyzed before, compared with P-MPLS, multiblock modeling allows for easier interpretation of both the roles of each local phase and the integrated contribution of all phases. Moreover, the phase behaviors are explored under the influence of covarying systematic dynamics among phases. A multiblock PLS-CCA method (MBPLS-CCA), which is shown in Appendix B, is developed here for between-phase cumulative analysis.

At the end of process, process information of all phases is available. Different phases, covering both critical and uncritical ones, may contribute to different parts of quality variations to different extents. Then their data blocks are well prepared in the common and uncommon subspaces as shown in Figure 1d:

$$\begin{aligned}\mathbf{X}_{\mathrm{MC}}(I \times \mathrm{JK}) &= \overset{\frown}{\mathbf{X}}_{\mathrm{MC}} + \overset{\smile}{\mathbf{X}}_{\mathrm{MC}} \\ \overset{\frown}{\mathbf{X}}_{\mathrm{MC}}(I \times \mathrm{JK}) &= \left[\overset{\frown}{\mathbf{X}}_1, \overset{\frown}{\mathbf{X}}_2, \ldots, \overset{\frown}{\mathbf{X}}_{\mathrm{C}}\right] \\ \overset{\smile}{\mathbf{X}}_{\mathrm{MC}}(I \times \mathrm{JK}) &= \left[\overset{\smile}{\mathbf{X}}_1, \overset{\smile}{\mathbf{X}}_2, \ldots, \overset{\smile}{\mathbf{X}}_{\mathrm{C}}\right]\end{aligned} \tag{13}$$

In different subspaces, by performing MBPLS-CCA modeling, the block information and super information are obtained and the end-of-process quality analysis results ($\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}$ and $\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}$) are obtained, respectively:

$$\begin{aligned}\hat{\overset{\frown}{\mathbf{X}}}_{MC,c} &= \overset{\frown}{\mathbf{T}}_{MC,c}\overset{\frown}{\mathbf{P}}{}^{\mathrm{T}}_{MC,c} \\ \hat{\overset{\smile}{\mathbf{X}}}_{MC,c} &= \overset{\smile}{\mathbf{T}}_{MC,c}\overset{\smile}{\mathbf{P}}{}^{\mathrm{T}}_{MC,c} \\ \hat{\overset{\frown}{\mathbf{Y}}}_{MC} &= \overset{\frown}{\mathbf{T}}_{T}\overset{\frown}{\mathbf{Q}}{}^{\mathrm{T}}_{MC} \\ \hat{\overset{\smile}{\mathbf{Y}}}_{MC} &= \overset{\smile}{\mathbf{T}}_{T}\overset{\smile}{\mathbf{Q}}{}^{\mathrm{T}}_{MC}\end{aligned} \tag{14}$$

where, $\overset{\frown}{\mathbf{T}}_{\mathrm{MC,c}}$ and $\overset{\smile}{\mathbf{T}}_{\mathrm{MC,c}}$ are the modeled block/phase scores in the common subspace and uncommon subspace, respectively. $\overset{\frown}{\mathbf{T}}_{\mathrm{T}}$ and $\overset{\smile}{\mathbf{T}}_{\mathrm{T}}$ are the super scores in two subspaces, respectively. $\overset{\frown}{\mathbf{P}}_{\mathrm{MC,c}}$ and $\overset{\smile}{\mathbf{P}}_{\mathrm{MC,c}}$ are loadings matrices for $\overset{\frown}{\mathbf{X}}_{\mathrm{MC,c}}$ and $\overset{\smile}{\mathbf{X}}_{\mathrm{MC,c}}$, respectively; $\overset{\frown}{\mathbf{Q}}_{\mathrm{MC}}$ and $\overset{\smile}{\mathbf{Q}}_{MC}$ are loadings matrices for qualities, respectively.

The final quality prediction at the end of each process can be also readily obtained by combining the predicted qualities from two different subspaces using weights derived from PLS-CCA algorithm:

$$\begin{aligned}&\left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}, \hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}\right] \xrightarrow{\mathrm{PLS-CCA}(\mathbf{W}_{y\mathrm{MC}})} \mathbf{Y} \\ &\hat{\mathbf{Y}}_{\mathrm{MC}} = \left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}, \hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}\right]\mathbf{W}_{y\mathrm{MC}} = \hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}\overset{\frown}{\mathbf{W}}_{y\mathrm{MC}} + \hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}\overset{\smile}{\mathbf{W}}_{y\mathrm{MC}}\end{aligned} \tag{15}$$

where, $\mathbf{W}_{\mathrm{yMC}}(4 \times 2) = \begin{bmatrix}\overset{\frown}{\mathbf{W}}_{\mathrm{yMC}} \\ \overset{\smile}{\mathbf{W}}_{\mathrm{yMC}}\end{bmatrix}$ is the weights corresponding to the two subspaces.

Comments and notes

Based on the above regression modeling and quality analysis strategy, not only the time-isolated effect but also both the within- and between-phase cumulative effects are explored. The separation of two different subspaces provides different analysis scenes which enclose different operation patterns and regression relationships. First, it improves the phase representability of the online quality analysis model by only focusing on the common patterns and excluding those time-varying patterns. Differently, the offline quality interpretation, covering both within- and between-phase cumulative analyses, is performed in both subspaces. The advantage of the subspace separation is that in addition to an analysis angle of view for the whole phase, one also obtains analysis angles for two different types of operation patterns. It is deemed that their respective contributions to qualities tend to be hidden by each other to a certain extent when the two subspaces are not separated. By zooming into separate subspace, it provides improved interpretation and potential for capturing the key factors in different subspaces for quality improvement. It is ready to know how the common and uncommon patterns will act under the influence of within- and between-phase correlations along time direction as well as their different influences on quality explanation. For example, during within-phase cumulative analysis, it is more general that the time-wise correlations may influence those uncommon patterns more seriously than common ones, which can be indicated by the regression weights attached to the regression variables, $\overset{\frown}{\mathbf{R}}_{\mathrm{c}}$ and $\overset{\smile}{\mathbf{R}}_{\mathrm{c}}$. In addition, by comparing $\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}$ and $\bar{\hat{\mathbf{Y}}}_{\mathrm{vc}}$ which is calculated by $\bar{\hat{\mathbf{Y}}}_{\mathrm{vc}} = \sum_{k\in c}\hat{\mathbf{Y}}_{k,\mathrm{vc}}$ and has been used as the end-of-phase quality prediction in the subPLS algorithm by Lu and Gao,[18] it may be found that they are quite different. It means that the simple average calculation of online quality predictions can not explore the within-phase cumulative effects. Moreover, comparing the modeled phase information in within- and between-phase cumulative analyses respectively, it can reveal the different influences of the between-phase correlations on the common and uncommon patterns. In summary, these modeling results provide a rational evaluation platform, where their modeled process variations and quality variations can then be quantitatively defined and then compared for an improved model interpretation and process understanding. This will be further illustratively analyzed and discussed in Simulation section.

It should be noted that the basic assumptions of our work cover two points: one is that part of covariances stay constant within the same phase, that is, one common subspace exists; the other is that there are critical-to-quality phases in which online quality prediction results can be reliably accepted. In some practical cases, this can be well satisfied, such as injection molding used in the work. However, for some cases, which are typically seen in fine chemicals and food processing, etc., the batches progress in an obvious cumulative manner. Under that situation, it may be impossible to find such a common subspace in which the **X–Y** covariances stay similar. Moreover, it is possible that each specific phase only makes a certain contribution to part of qualities, which means there is no critical phase in which accurate online quality predictions can be accepted satisfactorily. Fortunately, the two basic assumptions can be automatically checked in the current work. First, during the second-step phase division, whether the common subspace really exists

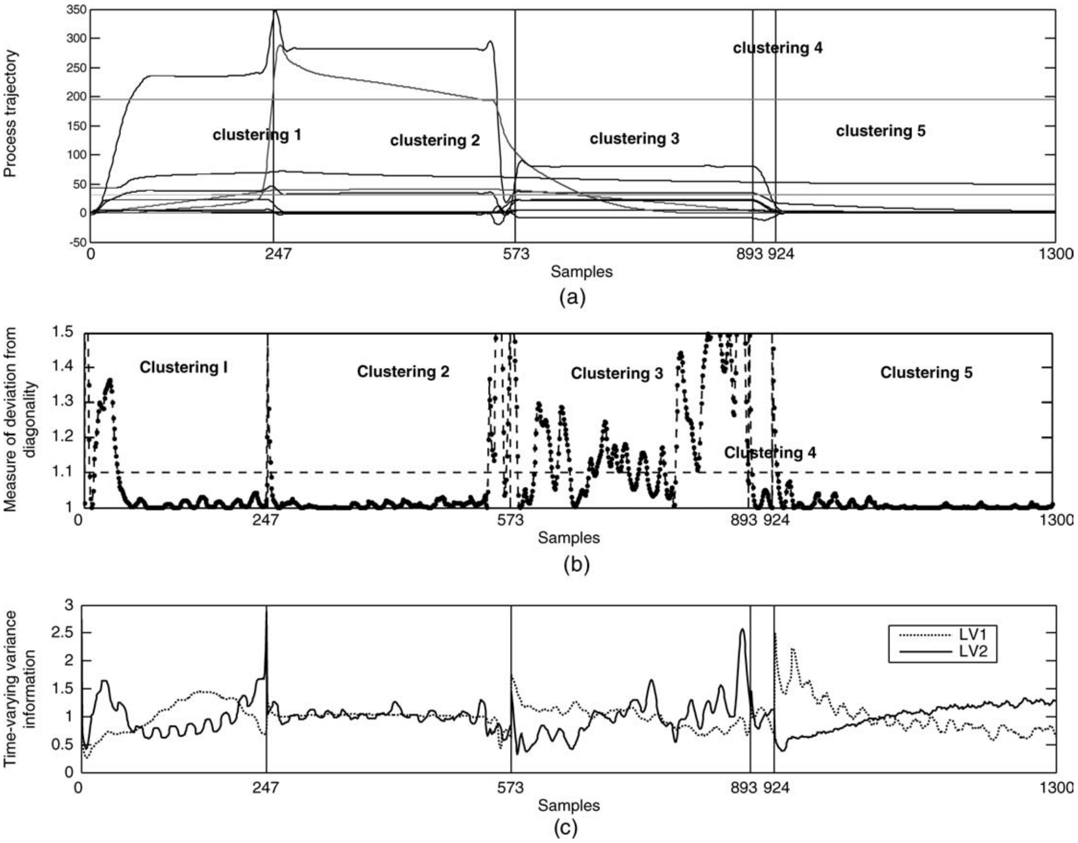


Figure 2. (a) Original subPLS clustering result and process trajectory; (b) evaluation result using measure of deviation from diagonality along time direction; (c) time-varying variance trajectories of the first two PLS-CCA LVs.

can be verified by the evaluation index defined in Eqs. 3 and 5. For the final phase division result, if $\overset{\frown}{A}_{\text{vc}}^{\text{pls}} = 0$, it means that there is no common subspace within each phase; If $\overset{\frown}{A}_{\text{vc}}^{\text{pls}} = 1$, then there is no need to perform CCA as postprocessing after PLS. The larger the dimension size $(\overset{\frown}{A}_{\text{vc}})$, the larger the PLS-CCA common subspace. Moreover, the critical phases can be checked by the prediction accuracy of quality variation during online prediction.

Moreover, it is worth remarking that the nature of the current modeling method is linear. That is, phase division is indicated by the changes of linear variable correlations and the phase representation is derived by linear transformations of the original data. Moreover, critical phases are also checked based on the quality prediction accuracy of linear models. However, for real industry batch processes, it is not uncommon that the underlying correlations are nonlinear to some extent, which introduces extra complications. For example, one critical nonlinear phase model may be fitted by multiple critical linear ones. Since linear statistical analysis techniques may not be competent enough to exploit the nonlinear data structure, it is desirable to employ nonlinear statistical analysis technique in phase division and model representation to handle the problem aroused by nonlinear behaviors. This is a meaningful issue and deserves further investigation. The current report provides the basis and potential for future work.

Simulations and Discussions

Process description

Injection molding,[38,39] 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 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. It is a typical MP batch process and has been widely used in previous work.[13,18,19] An injection molding machine is well instrumented in our lab and the authors have rich expertise knowledge of the injection molding. It can be readily implemented for experiments, in which, all key process conditions such as the temperatures, pressures, displacement, and velocity can be online measured by their corresponding transducers, providing abundant process information. It provides an ideal candidate for application and verification of the proposed phase-based process analysis and quality prediction algorithm.

The material used in this work is high-density polyethylene (HDPE). Twelve process variables are selected for modeling, which can be collected online from measurements with a set of sensors. Two-dimension indices, product length (mm) and weight (g), are chosen to evaluate the product

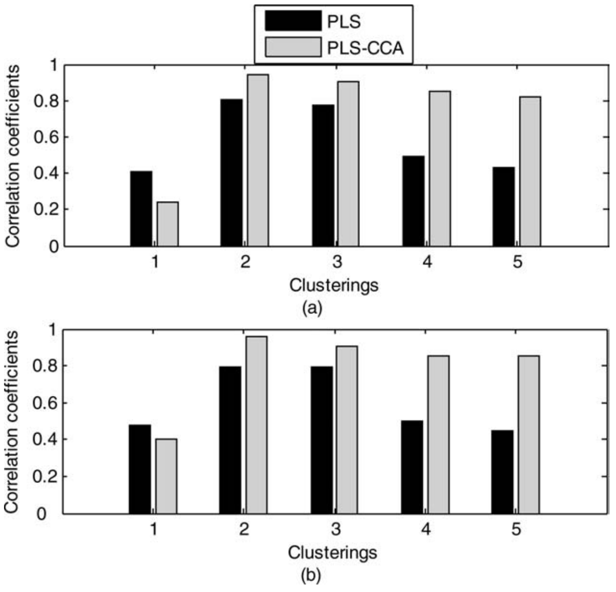


Figure 3. Correlation analysis between the first LV and quality variables (a) LV1-Y1 and (b) LV1-Y2.

quality, whose real values can be directly measured by instruments. Totally 50 normal batch runs are conducted under various operation conditions[18] by DOE method. Using injection stroke as indicator variable, the reference batches are unified to have even duration (1300 samples in this experiment) by data interpolation, which, thus, results in the descriptor array $\underline{\mathbf{X}}$ (50 × 12 × 1300). The qualities are only measured at the end of process, generating the dependent matrix $\mathbf{Y}$(50 × 2). The first 35 batches are used for modeling, while the other 15 cycles are used for model validation.

Phase division and subspace separation

First, 1300 normalized time-slices $\mathbf{X}_k$ $(I \times J_k)$ are obtained from $\underline{\mathbf{X}}$(35 × 12 × 1300) as well as the normalized quality variables $\mathbf{Y}$(35 × 2). Then 400 time-slice weight matrices are obtained focusing on the normalized data sets, $\{\mathbf{X}_k, \mathbf{Y}\}$. They are weighted using the time-varying variances of PLS LVs and then fed to the first-step clustering algorithm. The process duration is preparatorily partitioned into five main clusters, in which, operation time information is included so that process samples are consecutive within the same clustering. The original phase clustering result is shown in Figure 2a combined with the time-varying process operation trajectory. The first-step clustering result is deemed to be consistent with the real four physical operation phases: injection, packing-holding, plastication, and cooling. Moreover, the short period between Clusters 3 and 5, i.e., Cluster 4, may reveal the gradual transition when process operation is alternating from plastication to cooling.

Although the first-step clustering result seems to coincide with the real physical operation phases, which, however should be evaluated to check whether they are consistent with the changes of underlying characteristics and whether the resulting cluster model is representative enough. From the evaluation results shown in Figure 2b, it can be clearly seen that generally the operation patterns in every cluster can not be comprehensively described by a certain uniform phase model. Larger the deviation amplitude, less representative the cluster-based model. Here temporarily set 1.1 as the threshold. Those exceptional points between neighboring clusters may reveal the transition characteristics from one phase to another. Besides them, in the starting periods of both Clusters 1 and 5, the statistical index values are significantly larger than 1, and especially obvious for the entire Cluster 3, revealing that it is not proper to classify the corresponding patterns into the same modeling phase. That is, they can not be described by a unified phase model. Moreover, the variances of the extracted time-slice PLS-CCA LVs are also shown in Figure 2c, which shows a time-varying trend along time direction and well coincides with the previous statement after Eq. 6. Based on the evaluation result, the first-step clustering result should be further subdivided to derive real phase-representative models. Moreover, to reveal the effects of CCA post-processing following PLS, the correlation coefficients between the first LV, which is extracted by subPLS and PLS-CCA, respectively, and the two quality variables are calculated in each cluster. For the simplicity of comparison, their absolute values are used and contrastively shown in Figure 3. Using CCA post-processing, generally the relationships between the extracted LVs and qualities are enhanced, yielding higher correlation coefficients.

Then, using the second-step subdivision, the subspace separation and the final phase division results are obtained. As shown in Figure 4a, the quantitative evaluation analysis is performed focusing on the common subspace. Comparing the φ_k plot with that shown in Figure 2b, it is clear the phase representability in the common subspace has been greatly improved. Moreover, the variance trajectories are shown in Figure 4b. Here the minimum phase length is set 20. Finally, we can get four main longer phases (Phases 5, 7, 13, and 23), which are all longer than 100 sampling intervals and multiple shorter phases. Those transition patterns, although they are shorter than the predefined threshold (*minlenphase*), are also isolated from their neighboring phases so that they will not deteriorate the representability of phase models. Here for simplicity of designation, all subdivided time segments are uniformly called ''phase,'' in which, the transition regions, however, are not used for quality analysis. The formal phases, excluding those transition regions, are shown in Tables 1 and 2, in which, the quality analyses can be performed both online and offline.

Phase-based quality analysis

In the common subspace, the online quality prediction model is derived. The separated common process variations as well as the predicted quality variations at each time are quantitatively evaluated based on the training batches and shown in Figure 5. Here, for visual continuity, the analysis is performed at all time. Although the prediction results are not accurate enough in some regions such as transition ones, they can reveal the time-varying trend. From the plot, at different time especially over different phases the modeled process variations are significantly different. The two predicted quality variables show very similar variation trajectories throughout the process duration except the time region

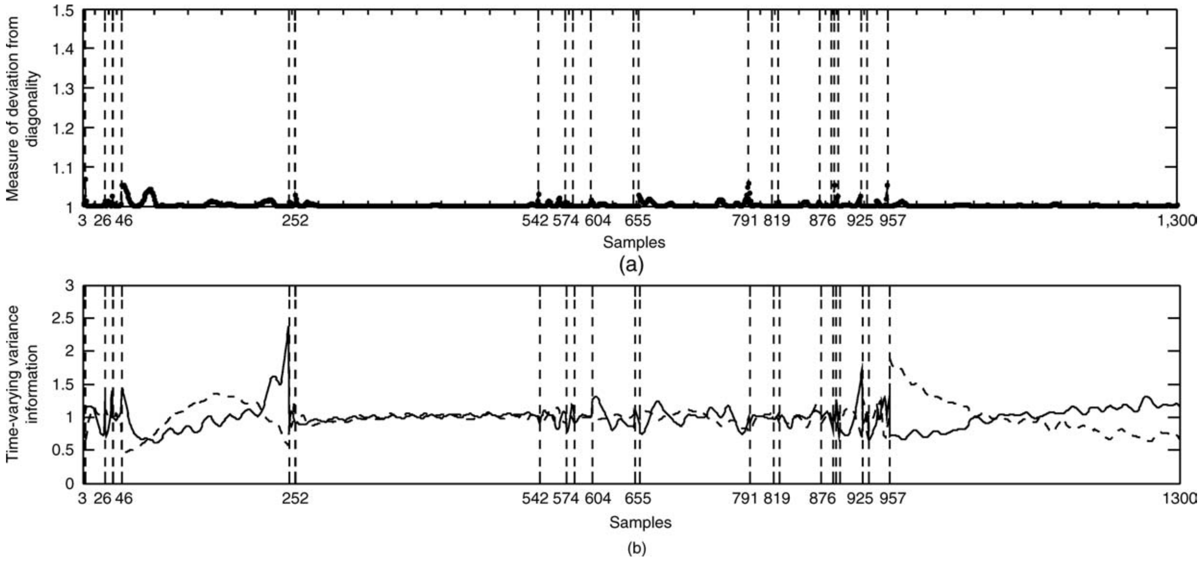


Figure 4. Second-step phase subdivision and evaluation results.

(a) Measure of deviation from diagonality along time direction and (b) time-varying variance trajectories of the first two PLS-CCA LVs.

spanning 46–245, i.e., Phase 5. To identify the critical phases, in which online predictions can be accepted with credibility, the goodness-of-fit of the phase models is measured by multiple coefficient of determination, $R^2_{j_y,c} = \frac{\sum_{k\in c}\sum_{i=1}^{I}\hat{y}^2_{i,k,j_y}}{\sum_{k\in c}\sum_{i=1}^{I}y^2_{i,k,j_y}}$, as done in our previous work.[18] Here, Phases 7, 8, 10, and 11 are judged to be critical ones by $R^2\hat{\mathbf{Y}}_{\mathrm{vc}}$ shown in Tables 1 and 2.

Then the within-phase cumulative analysis is carried out in both subspaces. Taking example for the first 10 time-slices in Phases 7 and 8, respectively, both of which span 120 process variables in all, the regression weights, $\overset{\frown}{\mathbf{R}}_{\mathrm{c}}$, in the common subspace are shown in Figure 6a in comparison with $\overset{\smile}{\mathbf{R}}_{\mathrm{c}}$ in the uncommon subspace shown in Figure 6b. It is observed that within the common subspace the weights have a certain similarity over time but with different magnitudes. In contrast, the weights are quite different over time in the uncommon subspace. It reveals that time correlations impose more serious influences on those uncommon patterns. Moreover, the joint end-of-phase predictions can be obtained by combining the values in two subspaces ($\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}$ and $\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{c}}$). Figures 7a shows the four weights attached to them for the two quality variables $\mathbf{Y}1$ and $\mathbf{Y}2$, respectively. The weights are quite different over phases, revealing their different contributions and phase-specific tokens. Besides the within-phase correlations, various phases may also correlate and influence with each other, which can be explored by multiblock modeling (MBPLS-CCA). Under the influence of between-phase covarying dynamics, the phase behaviors are modeled differently more or less. In two different subspaces, we can get two-dimensional quality predictions respectively, $\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}$ and $\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}$, Figure 7b shows the weights attached to the four regressors

Table 1. Modeled Process and Quality Variations During Online and Offline Analysis

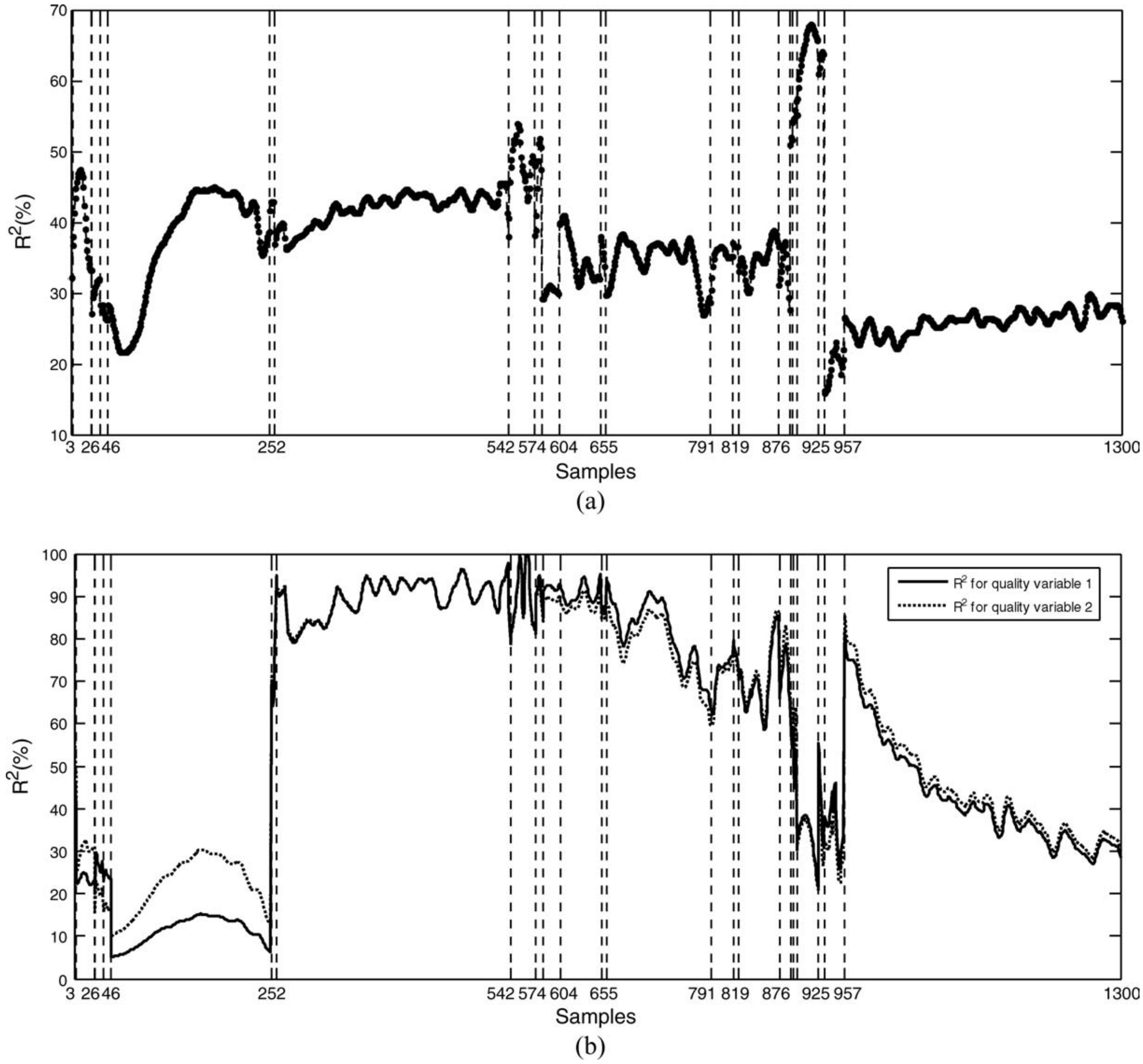
	Process Variations R^2 (%)					Predicted Quality Variations R^2 (%)									
	$R^2\overset{\frown}{\mathbf{X}}_{\mathrm{vc}}$	$R^2\hat{\overset{\frown}{\mathbf{X}}}_{\mathrm{c}}$	$R^2\hat{\overset{\smile}{\mathbf{X}}}_{\mathrm{c}}$	$R^2\hat{\overset{\frown}{\mathbf{X}}}_{\mathrm{MC,C}}$	$R^2\hat{\overset{\smile}{\mathbf{X}}}_{\mathrm{MC,C}}$	$R^2\hat{\mathbf{Y}}_{\mathrm{vc}}$		$R^2\hat{\bar{\mathbf{Y}}}_{\mathrm{vc}}$		$R^2\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}$		$R^2\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{c}}$		$R^2\hat{\mathbf{Y}}_{\mathrm{c}}$	
Phase no.	$\frac{\sum\overset{\frown}{\mathbf{X}}^2_{\mathrm{vc}}}{\sum\mathbf{X}^2_{\mathrm{vc}}}$	$\frac{\sum\hat{\overset{\frown}{\mathbf{X}}}^2_{\mathrm{c}}}{\sum\mathbf{X}^2_{\mathrm{c}}}$	$\frac{\sum\hat{\overset{\smile}{\mathbf{X}}}^2_{\mathrm{c}}}{\sum\mathbf{X}^2_{\mathrm{c}}}$	$\frac{\sum\hat{\overset{\frown}{\mathbf{X}}}^2_{\mathrm{MC,c}}}{\sum\mathbf{X}^2_{\mathrm{c}}}$	$\frac{\sum\hat{\overset{\smile}{\mathbf{X}}}^2_{\mathrm{MC,c}}}{\sum\mathbf{X}^2_{\mathrm{c}}}$	$\frac{\sum\hat{\bar{\mathbf{Y}}}^{*2}_{\mathrm{vc}}}{\sum\mathbf{Y}^2_{\mathrm{vc}}}$		$\frac{\sum\hat{\bar{\mathbf{Y}}}^2_{\mathrm{c}}}{\sum\mathbf{Y}^2}$		$\frac{\sum\hat{\overset{\frown}{\mathbf{Y}}}^2_{\mathrm{c}}}{\sum\mathbf{Y}^2}$		$\frac{\sum\hat{\overset{\smile}{\mathbf{Y}}}^2_{\mathrm{c}}}{\sum\mathbf{Y}^2}$		$\frac{\sum\hat{\mathbf{Y}}^2_{\mathrm{c}}}{\sum\mathbf{Y}^2}$	
2	41.39	30.09	8.31	27.69	10.66	23.41	30.07	22.60	29.05	35.25	40.52	70.56	74.78	70.61	74.80
5	37.07	24.87	24.48	26.33	25.13	11.22	22.42	10.47	20.92	79.41	81.65	93.04	90.51	93.10	91.25
7	42.06	38.58	8.60	41.18	9.16	89.61	89.86	88.95	89.21	95.76	96.42	98.29	97.30	98.29	97.56
8	48.48	45.45	9.14	47.42	12.85	90.26	89.71	88.46	87.92	95.17	96.71	95.45	96.77	95.73	97.16
10	30.34	26.85	16.47	29.52	40.58	92.02	89.26	91.78	89.02	93.50	91.42	78.24	87.96	93.52	94.14
11	35.04	26.66	3.85	29.08	5.71	90.52	88.01	89.50	87.14	94.06	94.93	96.44	97.91	96.68	98.04
13	34.67	25.95	7.80	30.99	5.81	81.51	77.83	78.82	75.46	95.33	94.02	98.53	97.93	98.54	97.93
14	35.13	24.27	12.05	29.40	7.06	71.42	69.98	67.80	66.51	88.56	83.40	89.56	94.06	93.70	94.88
16	34.85	26.77	17.35	32.73	37.23	69.75	71.25	66.94	68.45	87.23	89.61	98.14	95.62	98.24	96.61
20	64.77	36.16	10.78	56.33	10.48	33.65	32.67	30.43	29.55	39.75	37.93	0.50	1.50	40.61	38.16
22	19.63	14.25	32.25	14.44	45.57	36.33	31.22	33.01	28.99	51.04	43.68	79.52	78.00	79.66	78.13
23	26.08	21.72	23.12	24.93	34.57	42.75	44.91	37.00	38.87	81.20	86.59	97.90	96.02	97.99	96.49

Table 2. End-of-Process Modeled Quality Variations

R^2 (%)					
$R^2\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}$		$R^2\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}$		$R^2\hat{\mathbf{Y}}_{\mathrm{MC}}$	
$\frac{\sum \hat{\overset{\frown}{\mathbf{Y}}}{}^2_{\mathrm{MC}}}{\sum \mathbf{Y}^2}$		$\frac{\sum \hat{\overset{\smile}{\mathbf{Y}}}{}^2_{\mathrm{MC}}}{\sum \mathbf{Y}^2}$		$\frac{\sum \hat{\mathbf{Y}}^2_{\mathrm{MC}}}{\sum \mathbf{Y}^2}$	
99.77	96.69	97.15	98.82	99.72	99.09

enclosed in $\left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}\right]$ to get the joint end-of-process quality predictions. For example, corresponding to the first quality variable (**Y**1), weights to its predicted values in $\left[\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{MC}}\hat{\overset{\smile}{\mathbf{Y}}}_{\mathrm{MC}}\right]$ are significantly larger than those attached to the predicted values of **Y**2, which agrees well with the real case.

On the basis of phase-specific subspace separation, both online and offline quality analyses are performed with different focuses and application purposes. More details about quality-relevant underlying information can thus be obtained. The modeling result is summarized in Tables 1 and 2, where both process and quality variations are quantitatively evaluated and compared, revealing the changes of their roles and manners over phases. Especially, how the operation patterns in the two subspaces act in quality analysis under the influence of within- and between-phase time-wise covarying dynamics is also exposed. First, in the common subspace, the simple average calculation of all online quality predictions within the same phase may sacrifice some quality information, which can be clearly seen by comparing $\boldsymbol{R}^{\mathbf{2}}\bar{\hat{\hat{\mathbf{Y}}}}_{\mathrm{vc}}$ with $R^2\hat{\mathbf{Y}}_{\mathrm{vc}}$. Moreover, the modeled quality variations by the end-of-phase analysis ($\boldsymbol{R}^{\mathbf{2}}\hat{\overset{\frown}{\mathbf{Y}}}_{\mathrm{c}}$) which takes the within-phase time correlations into account are higher than those modeled by online prediction ($R^2\hat{\mathbf{Y}}_{\mathrm{vc}}$). It demonstrates the cumulative effects of each phase. Besides, in the uncommon subspace, considerable contributions to quality prediction are also



(a)

(b)

Figure 5. Modeled variation information along time direction in online quality analysis.

(a) Process variations and (b) quality variations.

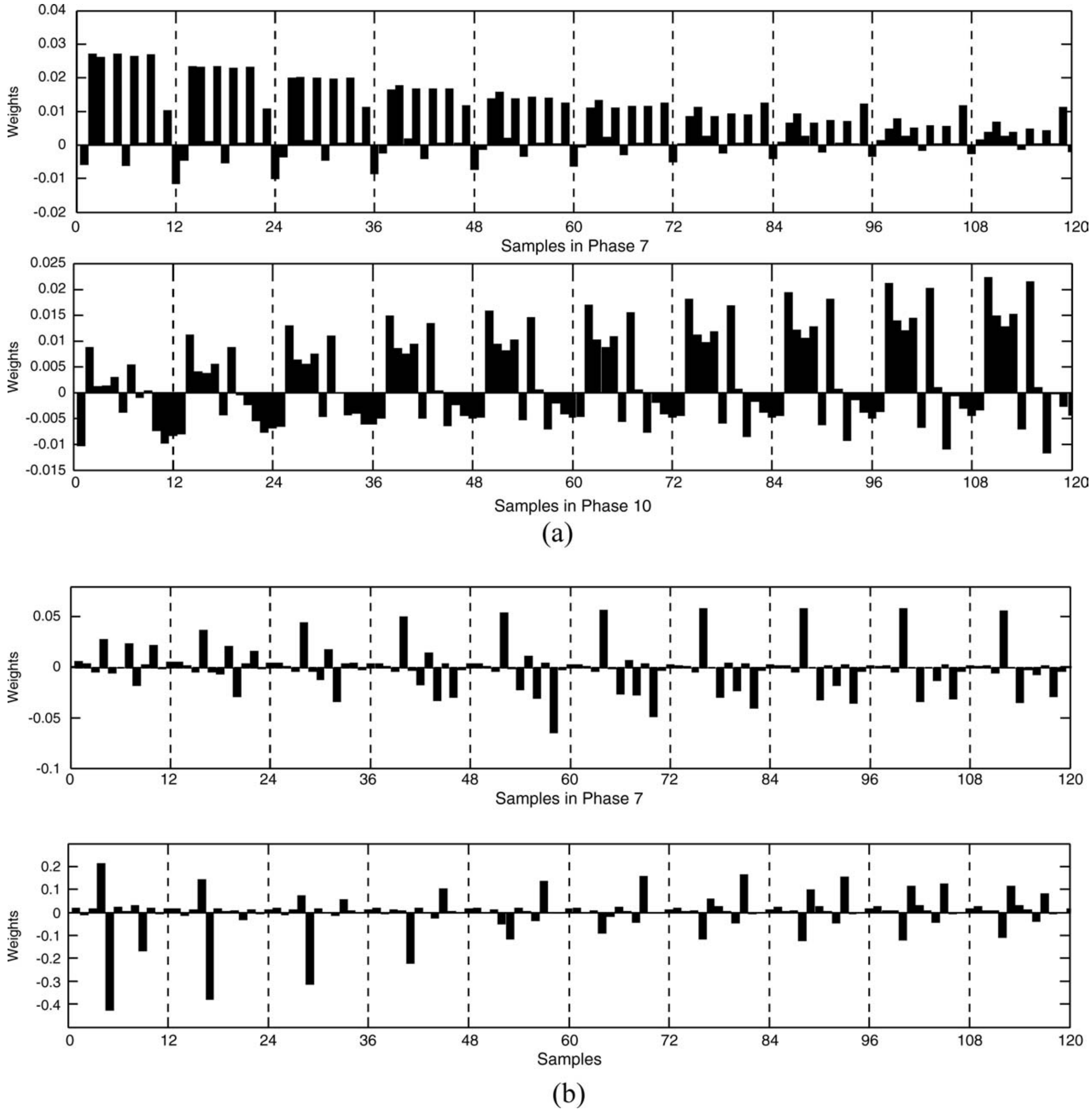
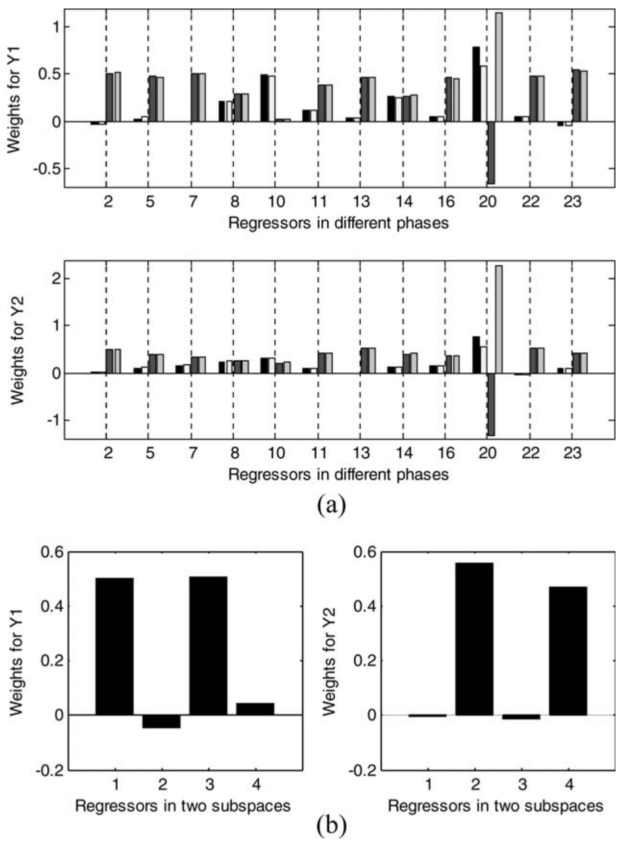


Figure 6. Regression weights to get the first LV in Phase 7 and Phase 10 during within-phase cumulative analysis (a) in the common subspace and (b) in the uncommon subspace.

revealed by $\boldsymbol{R}^2\hat{\overset{\smile}{\mathbf{Y}}}_\mathrm{c}$. Moreover, the joint end-of-phase quality variations ($R^2\hat{\mathbf{Y}}_\mathrm{c}$) are higher than both $\boldsymbol{R}^2\hat{\overset{\frown}{\mathbf{Y}}}_\mathrm{c}$ and $\boldsymbol{R}^2\hat{\overset{\smile}{\mathbf{Y}}}_\mathrm{c}$, which means that the two subspaces complementarily explain the qualities. However, they do not satisfy the simple sum relationship: $\boldsymbol{R}^2\hat{\mathbf{Y}}_\mathrm{c} \neq \boldsymbol{R}^2\hat{\overset{\frown}{\mathbf{Y}}}_\mathrm{c} + \boldsymbol{R}^2\hat{\overset{\smile}{\mathbf{Y}}}_\mathrm{c}$. This reveals that the modeled quality information in the two different subspaces may also overlap with each other. Besides, the modeled process variations reveal the details of the process behaviors in quality description over phases. For example, generally $\boldsymbol{R}^2\overset{\frown}{\mathbf{X}}_\mathrm{vc} > \boldsymbol{R}^2\hat{\overset{\frown}{\mathbf{X}}}_\mathrm{c}$, which indicates that when the common patterns are put together batch-unfolding for the end-of-phase quality interpretation, they are modeled in a different way from that when they are isolated for online application. Summing the process variations and quality variations modeled over all phases respectively, we get values much more than 100%. It demonstrates that various phases cover overlapping information resulting from the between-phase correlations. They are not explored by the end-of-phase analysis which considers each phase separately but will be modeled well by the end-of-process analysis in which the between-phase correlations are respected well. Therefore, the modeled process variations $\boldsymbol{R}^2\hat{\overset{\frown}{\mathbf{X}}}_{\mathrm{MC},\boldsymbol{c}}$ and $\boldsymbol{R}^2\hat{\overset{\smile}{\mathbf{X}}}_\mathrm{MC,c}$ are different from $\boldsymbol{R}^2\hat{\overset{\frown}{\mathbf{X}}}_\mathrm{c}$ and $\boldsymbol{R}^2\hat{\overset{\smile}{\mathbf{X}}}_\mathrm{c}$ more or less, in which, some phase information is hidden whereas some information is newly exposed under the



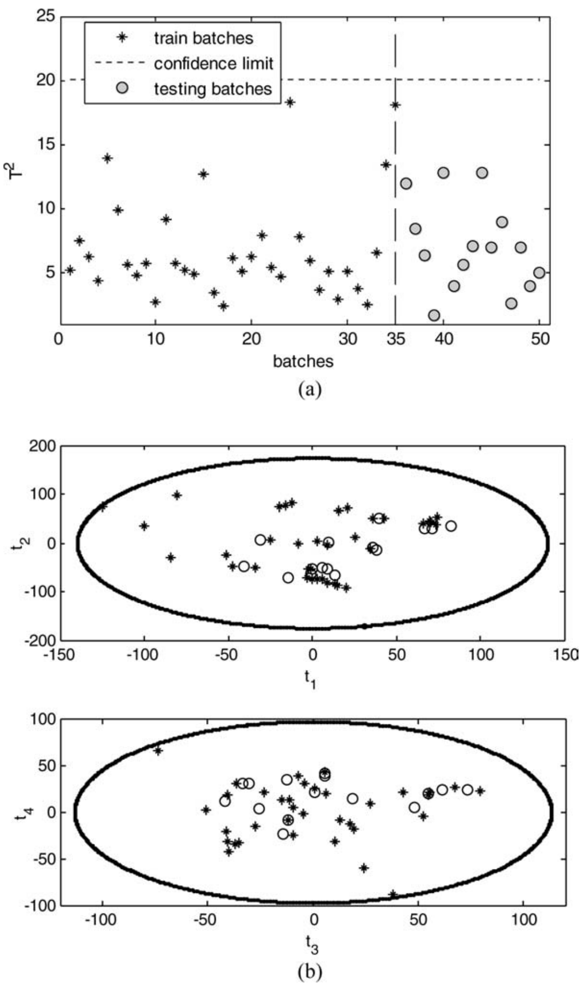
(a)

(b)

Figure 7. Regression weights attached to the four regressors for joint (a) within-phase and (b) between-phase quality analysis.

influence of between-phase correlations. It tells one when the between-phase correlations are modeled, the manner or contribution of various phases may be recomposed to a certain extent. Generally, the more serious the between-phase correlations are, the greater the recomposition is.

For the test batches, first a post analysis is performed to evaluate their similarity to the training batches. In Figure 8a, the test batches are projected onto a lower dimensional feature space enclosed by the conventional MPLS[7] on the training data and their variations are evaluated by the T^2 statistics. It is clear that they stay well within the normal region defined by the training batches. Further, to visually track the progress of their projections onto the feature space, the two-dimensional LV planes are shown in Figure 8b taking $\mathbf{t}_1$–$\mathbf{t}_2$ and $\mathbf{t}_3$–$\mathbf{t}_4$ planes, for example, which both reveal that the test batches are similar to some of the training ones. However, it does not mean that these test batches can get accurate quality predictions, which should depend on the fitting ability of the regression models to those similar training cycles. For the 15 test cycles, quality analysis results are simply summarized in Table 3. In those critical phases, the maximal and minimal values of online quality variations are calculated for both quality variables. And in all formal phases, the joint end-of-phase quality predictions are evaluated, as well as the final end-of-process ones, which, comparatively, are generally worse than those for training batches shown in Tables 1 and 2.



(a)

(b)

Figure 8. (a) T^2 plot for both training and test batches (b) t1–t2 and t3–t4 planes (dash contour, the control boundary corresponding to 5% significant level; "*", train batches and "O", testing batches).

Table 3. Online and Offline Quality Analysis Result for Test Batches

	Predicted Quality Variations R^2 (%)					
	$R^2\hat{\mathbf{Y}}_{\mathrm{k,vc}}$					
Phase no.	Min Value		Max Value		$R^2\hat{\mathbf{Y}}_{\mathrm{c}}$	
2	–				60.11	68.25
5					87.18	89.48
7	54.57	60.24	73.94	80.39	91.85	97.25
8	61.06	66.30	95.03	98.81	88.55	96.22
10	89.40	94.74	93.97	99.58	96.78	91.32
11	86.42	90.81	95.49	97.87	94.52	95.35
13	–				95.18	96.71
14					92.41	92.61
16					95.79	91.43
20					17.63	13.77
22					76.30	75.19
23					96.93	94.97
$R^2\hat{\mathbf{Y}}_{\mathrm{MC}}$			98.35 98.86			

Conclusion

In the present work, a methodology to gain process understanding and improve quality analysis by phase-based subspace separation and exploring how process variables affect product quality in two different subspaces has been designed. The contribution of this research is twofold. On the one hand, it can automatically check the time-wise similarity of the underlying characteristics within each phase. Then a more general assumption, the phase-specific "partial similarity," is proposed. It can extract those really phase-common patterns for regression modeling, which, thus, improves the phase representability of online prediction model. On the other hand, the different underlying characteristics of the common and uncommon patterns are further analyzed offline and their respective cumulative effects on qualities are explored under the influence of within- and between-phase correlations. In this way, one can better comprehend the underlying phase behaviors and their manner in quality prediction. Application to injection modeling case illustrates its effectiveness.

Acknowledgments

The work is supported in part by China National 973 program (2009CB320603) and National Natural Science Foundation of China (60774068).

Literature Cited

1. Martens H, Naes T. *Multivariate Calibration, 2nd ed.* Chichester: Wiley, 1994.
2. Burnham AJ, Viveros R, MacGregor JF. Frameworks for latent variable multivariate regression. *J Chemometrics*. 1996;10:31–45.
3. Cserhati T, Kosa A, Balogh S. Comparison of partial least-square method and canonical correlation analysis in a quantitative structure-retention relationship study. *J Biochem Biophys Methods*. 1998;36:131–141.
4. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst*. 2000;125:2125–2154.
5. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods, 3rd ed.* California: Wadsworth Publishing Co Inc, 2003.
6. Ergon R. Reduced PCR/PLSR models by subspace projections. *Chemometrics Intellig Lab Syst*. 2006;81:68–73.
7. Nomikos P, MacGregor JF. Multi-way partial least squares in monitoring batch processes. *Chemometrics Intellig Lab Syst*. 1995;30:97–108.
8. Camacho J, Pico J. Multi-phase principal component analysis for batch processes modelling. *Chemometrics Intellig Lab Syst*. 2006;81:127–136.
9. Camacho J, Pico J. Online monitoring of batch processes using multi-phase principal component analysis. *J Process Control*. 2006;16:1021–1035.
10. Camacho J, Pico J. Multi-phase analysis framework for handling batch processes data. *J Chemometrics*. 2008;22:632–643.
11. Liu JL, Wong DSH. Fault detection and classification for a two-stage batch process. *J Chemometrics*. 2008;22:385–398.
12. Lu NY, Gao FR, Yang Y, Wang FL. PCA-based modeling and on-line monitoring strategy for uneven-length batch processes. *Ind Eng Chem Res*. 2004;43:3343–3352.
13. Zhao CH, Wang FL, Gao FR, Lu NY, Jia MX. Adaptive monitoring method for batch processes based on phase dissimilarity updating with limited modeling data. *Ind Eng Chem Res*. 2007;46:4943–4953.
14. Zhao CH, Wang FL, Lu NY, Jia MX. Stage-based soft-transition multiple PCA modeling and on-line monitoring strategy for batch processes. *J Process Control*. 2007;17:728–741.
15. Zhao CH, Wang FL, Gao FR. Improved calibration investigation using phase-wise local and cumulative quality interpretation and prediction. *Chemometrics Intellig Lab Syst*. 2009;95:107–121.
16. Yao Y, Gao FR. A survey on multistage/multiphase statistical modeling methods for batch processes. *Annu Rev Control*. 2009;33:172–183.
17. Lu NY, Gao FR, Wang FL. Sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE J*. 2004;50:255–259.
18. Lu NY, Gao FR. Stage-based process analysis and quality prediction for batch processes. *Ind Eng Chem Res*. 2005;44:3547–3555.
19. Lu NY, Gao FR. Stage-based online quality control for batch processes. *Ind Eng Chem Res*. 2006;45:2272–2280.
20. Zhao CH, Wang FL, Mao ZZ, Lu NY, Jia MX. Improved batch process monitoring and quality prediction based on multiphase statistical analysis. *Ind Eng Chem Res*. 2008;47:835–849.
21. Undey C, Cinar A. Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Syst Mag*. 2002;22:40–52.
22. Macgregor JF, Jaeckle C, Kiparissides C, Koutoudi M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J*. 1994;40:826–838.
23. Kourti T, Nomikos P, Macgregor JF. Analysis, monitoring and fault-diagnosis of batch processes using multiblock and multiway PLS. *J Process Control*. 1995;5:277–284.
24. Westerhuis JA, Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemometrics*. 1998;12:301–321.
25. Qin SJ, Valle S, Piovoso MJ. On unifying multiblock analysis with application to decentralized process monitoring. *J Chemometrics*. 2001;15:715–742.
26. Lopes JA, Menezes JC, Westerhuis JA, Smilde AK. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnol Bioeng*. 2002;80:419–427.
27. Duchesne C, MacGregor JF. Multivariate analysis and optimization of process variable trajectories for batch processes. *Chemometrics Intellig Lab Syst*. 2000;51:125–137.
28. Walmsley AD. Improved variable selection procedure for multivariate linear regression. *Anal Chim Acta*. 1997;354:225–232.
29. Galvão RKH, Araújo MCU, Fragoso WD, Silva EC, José GE, Soares SFC, Paiva HM. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemometrics Intellig Lab Syst*. 2008;92:83–91.
30. Fearn T. On orthogonal signal correction. *Chemometrics Intellig Lab Syst*. 2000;50:47–52.
31. Westerhuis JA, De Jong S, Smilde AK. Direct orthogonal signal correction. *Chemometrics Intellig Lab Syst*. 2001;56:13–25.
32. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *J Chemometrics*. 2002;16:119–128.
33. Hardoon DR, Szedmak S, Taylor JS. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput*. 2004;16:2639–2664.
34. Yamamoto H, Yamaji H, Fukusaki E, Ohno H, Fukuda H. Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. *Biochem Eng J*. 2008;40:199–204.
35. Yu HL, MacGregor JF. Post processing methods (PLS-CCA): simple alternatives to preprocessing methods (OSC-PLS). *Chemometrics Intellig Lab Syst*. 2004;73:199–205.
36. Noble B, Daniel JW. *Applied Linear Algebra*. Englewood Cliffs, NJ: Prentice Hall, 1977.
37. Flury BN, Gautschi W. An algorithm for simultaneous orthogonal transformation of several positive definite symmetrical-matrices to nearly diagonal form. *Siam J Sci Stat Comput*. 1986;7:169–184.
38. Yang Y, Gao F. Cycle-to-cycle and within-cycle adaptive control of nozzle pressures during packing-holding for thermoplastic injection molding. *Polym Eng Sci*. 1999;39:2042–2064.
39. Yang Y, Gao F. Adaptive control of the filling velocity of thermoplastics injection molding. *Control Eng Pract*. 2000;8:1285–1296.
40. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed. China Machine Press: Beijing, 2003.
41. Kutner Michael, Nachtsheim CJ, Neter J. *Applied Linear Regression Models*, 4th ed. Higher Education Press: Beijing, 2005.

Appendix A: The second-step phase subdivision strategy

The second-step subdivision strategy is implemented on the basis of the first-step clustering result to judge whether the clustering result is representative enough for each phase and correct them where necessary. Here two statistical indices are combined, the index φ defined in Eqs. 3 and 5 for measure of deviation from diagonality and the mean-squared prediction errors (MSPE), which is calculated using test data as $\mathrm{MSPE} = \frac{1}{I_{\mathrm{te}} J_y} \sum_{i=1}^{I_{\mathrm{te}}} \sum_{j=1}^{J_y} (\mathrm{y}_{i,jy} - \hat{\mathrm{y}}_{i,jy})^2$ (where, I_{te} is the number of test batches). They are used to check whether a common subspace can be derived and whether the subdivision can improve the performance of quality prediction. It should be pointed out that the second-step phase subdivision strategy can be readily extended to the cases using other multivariate calibration algorithms besides PLS-CCA algorithm used here.

The input of the second-step subdivision algorithm involves:

(a) One data set within the same cluster, which has been arranged variable-wise

(b) The minimum length of a phase (*minlenphase*)

(c) The respective improvement thresholds of φ and MSPE to accept a subdivision

(d) The maximum number of phases within each clustering (*maxnumphase*)

The output is the phase subdivision result for the current clustering.

The recursive calculation procedure is listed as below:

(1) Set the initial number of phases within this cluster *numphase* = 1.

(2) Input the prepared data set, perform PLS-CCA on them and figure out the possible common PLS-CCA subspace, in which, an interim $\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc0}}$-dimensional PLS subspace and a $\overset{\frown}{A}_{\mathrm{vc0}}$-dimensional CCA subspace are determined by cross-validation to achieve the least statistical value φ_0. Then calculate the resulting MSPE_0 based on the test data.

(3) For each sampling time (k) within the current data set, if the subdivision in k generates two segments of length higher than the minimum phase duration (*minlenphase*), perform PLS-CCA on the two segments, respectively; calculate the resulting $\varphi_1^{k\_1}$ and $\varphi_1^{k\_2}$ for both segments based on the same $\overset{\frown}{A}_{\mathrm{vc0}}$-dimensional PLS-CCA subspace (as well as the same interim $\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc0}}$-dimensional PLS subspace). Then calculate the corresponding quality prediction results $\mathrm{MSPE}_1^{k\_1}$ and $\mathrm{MSPE}_1^{k\_2}$ based on test data.

(4) Find the sampling time ($k^\bullet$) at which the average of $\varphi_1^{k^\bullet\_1}$ and $\varphi_1^{k^\bullet\_2}$ is lowest. Comparing both $\varphi_1^{k^\bullet\_1}$ and $\varphi_1^{k^\bullet\_2}$ with φ_0, if the improvement for either segment, $\varphi_1^{k^\bullet\_1} - \varphi_0$ and $\varphi_1^{k^\bullet\_2} - \varphi_0$, does not reach the predefined threshold, then stop this branch.

(5) Otherwise, comparing both $\mathrm{MSPE}_1^{k^\bullet\_1}$ and $\mathrm{MSPE}_1^{k^\bullet\_2}$ with MSPE_0, if the improvement for either segment, $\mathrm{MSPE}_1^{k^\bullet\_1} - \mathrm{MSPE}_0$ and $\mathrm{MSPE}_1^{k^\bullet\_2} - \mathrm{MSPE}_0$, does not reach the predefined threshold, then stop this branch.

(6) Otherwise, accept subdivision and update the number of phases in the initial cluster *numphase*: = *numphase* + 1. If the current *numphase* is no longer less than the predefined *maxnumphase*, stop the iteration procedure, output the subdivision result.

(7) Otherwise, recursively repeat steps (2)–(6) for either of the two resulting segments respectively, each now employed as the new input data set in step (2) as well as the updated initial parameters, $\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc0}}$, $\overset{\frown}{A}_{\mathrm{vc0}}$, φ_0, and MSPE_0.

Some discussions

The basic assumption of the two-step phase division strategy is that one common subspace really exists in each properly separated phase. Moreover, to guarantee that CCA can be performed as postprocessing after PLS, generally, the final dimensions of PLS common subspace ($\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}}$) and PLS-CCA common subspace ($\overset{\frown}{A}_{\mathrm{vc}}$) are expected to satisfy $2 \le \overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}} \le J$ and $1 \le \overset{\frown}{A}_{\mathrm{vc}} \le \min(\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}}, J_y)$ respectively. The final common subspace is restricted by $\min(\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}}, J_y)$ resulting from the CCA algorithm itself. Since in most cases, $\overset{\frown}{A}{}^{\mathrm{pls}}_{\mathrm{vc}} > J_y$, due to the contaminated PLS LVs by the quality-uninformative process variations, the final common subspace is actually directly confined by the number of quality variables. Therefore, a common subspace with at least two common directions can be figured out. However, sometimes it is possible that only one common direction exists ($\overset{\frown}{A}_{\mathrm{vc}} = 1$), which should be checked using another method different from the present one. Here, it is just simply described. Considering the sole common direction should have a direct and consistent regression relationship with the quality variables, the squared correlation coefficient between the time-varying first LV and qualities is used to evaluate the relationship, $\mathrm{corr}_{\mathrm{c}}^2 = \frac{1}{K_c \cdot J_y} \sum_{k \in c, j_y \in J_y} \mathrm{corr}_{\mathrm{k}}^2 \left(\mathbf{t}_{k,\mathrm{c}}, \mathbf{y}_{j_y} \right)$ (where, subscripts k, c, and J_y denote sampling time, phase, and quality variable, respectively). The larger the value of $\mathrm{corr}_{\mathrm{c}}^2$, the more similar the time-varying first LV will be to quality variables. It is well known that the squared correlation coefficient is actually the coefficient of determination in simple linear regression analysis.[40,41] Therefore, F-test[40,41] can be used here to define its critical value $\frac{\mathrm{corr_c}^2}{(1-\mathrm{corr_c}^2)/(I-2)} \sim F_\alpha(\overset{\frown}{A}_{\mathrm{vc}}, I - \overset{\frown}{A}_{\mathrm{vc}} - 1)$ (where, α is the significance level), which can determine whether it can be accepted. If it is declined, then no common subspace can be found out. In contrast, a one-dimensional common subspace can be figured out.

On the other hand, one restriction about the minimum phase length has been imposed so that only the time regions that are longer than the predefined threshold can be separated as phases. However, for those operation patterns between two neighboring phases, it is very likely that they have the transition characteristics,[14] showing the gradual changes of underlying correlations from one phase to another driven by the mechanical and physical operation principle. Those transition patterns may always show high deviation from diagonality (φ_k) whenever they are classified into any one phase. They should be isolated as individuals although they are often shorter than the predefined minimum phase length. This can be achieved by performing a post-processing procedure on the second-step phase subdivision result by checking the time-varying φ_k profile. Here for simplicity, they are also called phase but do not participate in quality analysis since they do not have stable regression relationships.

Appendix B: MBPLS-CCA Algorithm

Input data $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_B]$ and $\mathbf{Y}$

Step 1

Perform regular PLS-CCA on $\mathbf{X}$ and $\mathbf{Y}$ to obtain the original super scores $\mathbf{t}_\mathrm{T}$ and $\mathbf{u}$.

Step 2

Performing PLS-CCA between $\mathbf{X}_\mathrm{b}$ and $\mathbf{u}$,

Block weights:

$\mathbf{w}_\mathrm{b} = \mathbf{W}_\mathrm{b,pls} \cdot \mathbf{w}_\mathrm{b,cca}$ (where $\mathbf{W}_\mathrm{b,pls}$ is a $J_\mathrm{b} \times A_\mathrm{b,pls}$-dimensional PLS weights matrix, and $\mathbf{w}_\mathrm{b,cca}$ is a $A_\mathrm{b,pls}$-dimensional CCA weight vector; J_b is the number of process variables in block data $\mathbf{X}_\mathrm{b}$ and $A_\mathrm{b,pls}$ is the number of retained block PLS LVs.)

Block scores:

$\mathbf{t}_\mathrm{b} = \mathbf{X}_\mathrm{b}\, \mathbf{w}_\mathrm{b}$ and $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_B]$

Performing PLS-CCA between $\mathbf{T}$ and $\mathbf{u}$,

Super weights:

$\mathbf{w}_\mathrm{T} = \mathbf{W}_\mathrm{T,pls} \cdot \mathbf{w}_\mathrm{T,cca}$ (where $\mathbf{W}_\mathrm{T,pls}$ is a $B \times A_\mathrm{T,pls}$-dimensional PLS weights matrix, and $\mathbf{w}_\mathrm{T,cca}$ is a $A_\mathrm{T,pls}$-dimensional weight vector; $A_\mathrm{T,pls}$ is the number of retained super PLS LVs.)

Super scores:

$\mathbf{t}_\mathrm{T} = T \cdot \mathbf{w}_\mathrm{T}$

And get the new super quality score

$\mathbf{u}$: $\mathbf{q} = \mathbf{Y}^\mathrm{T}\, \mathbf{t}_\mathrm{T}/\mathbf{t}_\mathrm{T}^\mathrm{T}\, \mathbf{t}_\mathrm{T}$, $\mathbf{u} = \mathbf{Yq}/\mathbf{q}^\mathrm{T}\, \mathbf{q}$.

Compare the obtained new super score $\mathbf{t}_\mathrm{T}$ with the one in the preceding iteration step. If they are approximately equal ($\|\mathbf{t}_\mathrm{T,new} - \mathbf{t}_\mathrm{T,old}\| \le \varepsilon$), stop; else substitute the new $\mathbf{u}$ for the old one and begin another loop.

Step 3

Deflate residuals:

$$\mathbf{p}_\mathrm{b} = \mathbf{X}_\mathrm{b}^\mathrm{T}\mathbf{t}_\mathrm{b} \Big/ \mathbf{t}_\mathrm{b}^\mathrm{T}\mathbf{t}_\mathrm{b}$$

$$\mathbf{E}_\mathrm{b} = \mathbf{X}_\mathrm{b} - \mathbf{t}_\mathrm{b}\mathbf{p}_\mathrm{b}^\mathrm{T}, \quad \mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, \ \ldots\ \mathbf{E}_B]$$
$$\mathbf{F} = \mathbf{Y} - \mathbf{t}_\mathrm{T}\mathbf{q}^\mathrm{T}$$

From here, one can go to Step 1 to implement the above procedures for the new LVs, where $\mathbf{X}$ and $\mathbf{Y}$ are both replaced by their corresponding residual matrices $\mathbf{E}$ and $\mathbf{F}$.

Here some important points should be noted:

(a) In this algorithm, we employ regular PLS-CCA LVs obtained from the whole process data as the initial points as shown in Step 1. This usually makes the algorithm converge very quickly, which has been verified in the implementation procedure.

During the algorithm, to calculate each PLS-CCA block score and super score respectively, multiple PLS scores are prepared as shown in Step 2. This results from the consideration that PLS LVs prepare possible quality-related process variations for the following CCA postprocessing and then only those closely quality-related process information will be extracted from them by CCA.

By this algorithm, between-phase cumulative effects can be more clearly revealed. Multiple block scores, $\overset{\frown}{\mathbf{T}}_\mathrm{MC} = \left[\overset{\frown}{\mathbf{T}}_\mathrm{MC,1}, \overset{\frown}{\mathbf{T}}_\mathrm{MC,2}, \ldots, \overset{\frown}{\mathbf{T}}_\mathrm{MC,C}\right]$ and $\overset{\smile}{\mathbf{T}}_\mathrm{MC} = \left[\overset{\smile}{\mathbf{T}}_\mathrm{MC,1}, \overset{\smile}{\mathbf{T}}_\mathrm{MC,2}, \ldots, \overset{\smile}{\mathbf{T}}_\mathrm{MC,C}\right]$, are prepared in two different subspaces respectively, each block score in either subspace revealing their different roles and abilities in quality description under the influence of other blocks. From them, the super scores, $\overset{\frown}{\mathbf{T}}_\mathrm{T}$ and $\overset{\smile}{\mathbf{T}}_\mathrm{T}$, are retrieved corresponding to two subspaces respectively.

Manuscript received July 6, 2009, and revision received Feb. 26, 2010.