Phase-Based Joint Modeling and Spectroscopy Analysis for Batch Processes Monitoring

ARTICLE *in* INDUSTRIAL & ENGINEERING CHEMISTRY RESEARCH · DECEMBER 2009

Impact Factor: 2.59 · DOI: 10.1021/ie9005996

3 AUTHORS, INCLUDING:

Chunhui Zhao
Zhejiang University
91 PUBLICATIONS **580** CITATIONS
SEE PROFILE

Furong Gao
The Hong Kong University of Science and T…
267 PUBLICATIONS **3,770** CITATIONS
SEE PROFILE

Phase-Based Joint Modeling and Spectroscopy Analysis for Batch Processes Monitoring

Chunhui Zhao,[†] Furong Gao,*[,†] and Fuli Wang[‡]

Department of Chemical and Biomolecular Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, and College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, P.R. China

Spectroscopy is a useful tool for analyzing chemical information in batch processes. However, conventional spectroscopic techniques can lead to complex model structures and difficult-to-interpret results because of the direct use of redundant wavelength variables. To solve this problem, the ICA technique has been used first to reveal the underlying independent sources from the observed mixtures and their mixing coefficients. This analysis associates the constituent species with their respective effects on the mixture spectra, which could make more sense for spectroscopy. In the following article, the use of mixing coefficients instead of redundant wavelength variables can provide a convenient modeling platform from which the phase-specific characteristics of chemical reactions are readily identified. Consequently, a phase-based joint modeling framework is formulated for process monitoring by combining different latent-variable- (LV-) based algorithms. It decomposes different types of systematic variations in spectral measurements (**X**) and then sets up different monitoring systems for them. Monitoring different parts of **X** can provide abundant process information about the status of the operation from a comprehensive viewpoint, thus benefitting process understanding and fault detection. The effectiveness of this approach is demonstrated by application to two case studies.

1. Introduction

Batch processes play an important role in producing high-value-added chemical and biochemical products. Their physical states can be derived by conducting multivariate statistical analyses[1−7] on engineering process measurements, such as temperatures, pressures, and flow rates, at regular time intervals. Recently, increasing attention has been paid to spectroscopic analysis for online monitoring of chemical batch reactions.[8−16] Calibration models have also been built using spectroscopic data[17−19] to estimate reaction compound concentration or other material properties such as conversion, viscosity, and so on. Spectroscopic analysis has the advantage of being fast, robust, and nondestructive, and thus, spectroscopic techniques can provide a good alternative to classical process analytical methods.[12] Furthermore, specific chemical properties, such as the concentrations of spectroscopically active chemical species present in the reaction mixture, changes in solvent conditions, and the presence of spectroscopically active impurities, are made available. Such real-time and high-quality chemical information can reveal the actual variations occurring in the course of the reaction and check whether the process will remain in control. This is a major advantage compared to standard process measurements of physical conditions. An extensive list of online spectroscopic monitoring can be found in the review by Workman et al.[9] Moreover, Westerhuis et al.[12] discussed the role of spectroscopic analysis for batch process monitoring in terms of both current and potential advances.

Generally, spectroscopic data are considered as a particular form of process data for which the process variables are a series of spectral absorbances over a range of wavelengths. Accordingly, the philosophy of spectroscopic analysis for chemical batch monitoring is similar to that of traditional engineering analytical methods, where the reference behaviors of systematic variations are modeled and designed based on historical data of normal batches, and then the new batch operation is compared against these established references. This is commonly achieved by a conventional spectroscopy technique (multiway principal component analysis, MPCA)[11,12] that directly decomposes the correlation among spectral variables and deals with the redundancy problem by feature extraction. It results from the consideration that variable collinearity is typical in spectral absorbance over a range of wavelengths. It thus allows the original spectral data space to be shrunk into a lower-dimensional feature subspace, and then employs those underlying features for further modeling. Moreover, if one wants to monitor the variations in the process variables that are most influential on the quality variables (**Y**), one can perform multiway partial least-squares (MPLS)[20] decomposition on **X** and design the monitoring system in a manner similar to that for MPCA. However, because there are hundreds of wavelength variables or even more in spectral measurements, generally, many modeling features are required, which makes the monitoring model difficult to interpret. Moreover, when it refers to the description of process dynamics, lagged wavelength variables are introduced and incorporated with the current variables to construct an expanded modeling unit, which can readily lead to a further significant increase of the number of variables and make the model even harder to interpret. Although the redundant problem of measurement variables has been analyzed and solved in engineering process analysis, spectroscopy has its own characteristics that require further consideration.

Actually, a spectral measurement of a mixture is often a linear combination of the spectra of the constituent species. It would be useful if the component spectra could be recovered from the mixture spectra.[21−23] The goal is the assessment and resolution of all the possible conformations, including intermediate states, and the estimation of their pure spectra and concentration profiles changing throughout the chemical reaction, which are all closely related to the end-product quality. However, neither

* To whom correspondence should be addressed. Tel.: +852-23587136. Fax: +852-23580054. E-mail: kefgao@ust.hk.
[†] The Hong Kong University of Science and Technology.
[‡] Northeastern University.

10.1021/ie9005996
Published on Web 12/03/2009

MPCA[11,12] nor MPLS[20] can achieve this purpose. Recently, multivariate curve resolution (MCR)[24−26] analysis and the independent component analysis (ICA) algorithm[21−23] have been developed, which both allow for the resolution of concentration profiles and pure spectra of different species from spectral measurements but using different calculation procedures. For MCR analysis, an algorithm based on alternating least-squares (MCR-ALS)[25,26] has been widely used, which is a two-step factor analysis process. Matrix decomposition is used as an initial step to generate abstract factors and scores. These initial estimates are then refined in a second-step optimization to produce pure-component spectra and an associated concentration profile. This second step involves a rotation, which utilizes a least-squares fit between the initial factor estimates and a target matrix. ICA[27−29] has shown that it is not generally sufficient to consider only up to second-order content (correlations, covariances) when describing the measured variables. It attempts to recover statistically independent signal sources instead of merely uncorrelated ones, given only observations that are assumed to be linear mixtures of the original signals. To achieve this end, ICA makes full use of higher-order statistics and applies criteria related to information theory and entropy. In probability theory, independence is a high-order statistic that guarantees uncorrelatedness, so that it is a much stronger condition than uncorrelatedness. It is designed based on higher-order statistics to extract the source species in the form of independent components (ICs) and to determine their effects on the observed mixture spectra, which is called blind source signal separation process. Previous works[21−26] have applied MCR or ICA to spectral data, successfully demonstrating the effectiveness of these approaches in recovering the components of interest from spectral mixtures and estimating their concentrations. However, considering the complexity of practical chemical processes, the foregoing reviewed works have not provided a comprehensive analysis of underlying chemical characteristics referring to the spectroscopic monitoring issue:

(1) Because the chemical characteristics are time-varying in a batch reaction, a uniform statistical model throughout the batch duration might not correctly track the changes of the underlying reaction status through time. In particular, for multiphase chemical reaction processes, the exploration of phase-specific underlying chemical characteristics might be more desirable, which, however, has not been adequately addressed.

(2) Because of the complexity of chemical reaction, different types of variation characteristics exist in spectral measurements that are differently influential on the end product. The real chemical reaction status can be more efficiently described and more chemical reaction information can also be revealed if these variations can be captured and monitored. This issue, however, has not yet been brought into focus.

Based on the above analyses, a phase-based spectroscopy analysis strategy is presented for improving process understanding and monitoring. It will design a monitoring system by addressing the following problems given a set of spectra collected along different runs of a chemical reaction: (a) How many chemical constituents are responsible for the changes in the observed mixture spectra? (b) What are the roles of constituent spectral features? (c) How do the concentrations of these constituent species change, showing both within-batch and across-batch dynamics? (d) How can the estimated species and their concentrations be used to design the monitoring system? To achieve the above goal, first, considering its specific effect on spectroscopy, ICA is used to separate both constituent spectra and their mixing relationships from mixture spectra, which respectively reveal different chemical reactants occurring throughout different reaction steps and their concentrations (i.e., their direct contributions to spectral measurements). Then, the mixing relationship is utilized directly as the modeling unit instead of the original spectral wavelengths. This provides a convenient analysis platform according to which the following modeling strategy can be readily implemented, involving the identification of multiple phases, consideration of timewise dynamics, and discrimination of various variations according to their direct influence on product quality. The resulting joint monitoring systems can thus reveal more comprehensive variation information that is beneficial to fault detection.

This article is organized as follows: In the next section, the knowledge of ICA is introduced, including its basic theory and model representation. The phase-based joint statistical modeling method is then described in detail and its underlying principles are also explained. In section 3, the application of the proposed method to two case studies demonstrates its effectiveness. Results are presented and discussed. Finally, conclusions are drawn in the last section.

2. Methodology

2.1. ICA Algorithm. Independent component analysis (ICA)[27−29] is a recently developed method in which the goal is to find the statistically independent, or as independent as possible, non-Gaussian hidden factors that constitute the observed variables through linear combinations. Such a representation can capture the essential structure of the measurement data in many applications, including feature extraction and signal separation.

Assuming that the J measured variables $x_1, x_2, ..., x_J$, can be described as a linear combination of R (generally $R \leq J$) non-Gaussian independent components $s_1, s_2, ..., s_R$, the basic matter of ICA is to estimate the latent components $\mathbf{s}$ and the mixing and demixing relationships, $\mathbf{A_x}$ ($R \times J$) and $\mathbf{W_x}$ ($J \times R$), from only the process measurements $\mathbf{x}$ without any related prior knowledge, called the process of blind separation. Here, "blind" means that one knows very little, if anything, about the mixing matrix and makes few assumptions about the source signals. Their relationship can be expressed as

$$\begin{aligned} \mathbf{s}^{\mathrm{T}} &= \hat{\mathbf{x}}^{\mathrm{T}}\mathbf{W_x} \\ \hat{\mathbf{x}}^{\mathrm{T}} &= \mathbf{s}^{\mathrm{T}}\mathbf{A_x} \end{aligned} \tag{1}$$

where $\mathbf{s}$ ($R \times 1$) denotes the independent component vector, which has unit variance: $E(\mathbf{ss}^{\mathrm{T}}) = \mathbf{I}$.

Before applying an ICA algorithm, it is assumed that the process variables have been prewhitened, which is generally achieved by PCA, so that the components are uncorrelated and their variances are equal to unity. This whitening transformation can be expressed as

$$\mathbf{z} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{x} = \mathbf{Qx} \tag{2}$$

where $\mathbf{Q} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^{\mathrm{T}}$ is the whitening matrix, in which $\mathbf{U}$ (orthogonal matrix of eigenvectors) and $\mathbf{\Lambda}$ (diagonal matrix of eigenvalues) are generated from the eigenvalue decomposition of the covariance matrix $E(\mathbf{xx}^{\mathrm{T}})$.

Naturally, $\mathbf{s}$ can be estimated and the process information can be reconstructed by

$$\begin{aligned} \mathbf{s}^{\mathrm{T}} &= \mathbf{z}^{\mathrm{T}}\mathbf{W}_{\mathrm{z}} \\ \hat{\mathbf{z}}^{\mathrm{T}} &= \mathbf{s}^{\mathrm{T}}\mathbf{A}_{\mathrm{z}} \end{aligned} \tag{3}$$

where $\mathbf{W}_{\mathrm{z}}$ is an orthogonal matrix, given that $E(\mathbf{ss}^{\mathrm{T}}) = \mathbf{W}_{\mathrm{z}}{}^{\mathrm{T}}E(\mathbf{zz}^{\mathrm{T}})\mathbf{W}_{\mathrm{z}} = \mathbf{W}_{\mathrm{z}}{}^{\mathrm{T}}\mathbf{W}_{\mathrm{z}} = \mathbf{I}$. Moreover, in ICA, the following

relationship always applies: $\mathbf{A}_z^T = \mathbf{W}_z$, that is, $\mathbf{a}_z\ (N \times 1) = \mathbf{w}_z\ (N \times 1)$. Thus, $\mathbf{A}_z\mathbf{A}_z^T = \mathbf{I}\ (R \times R)$, resulting from the calculation

$$\begin{aligned} \mathbf{W}_z &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{S} \\ \mathbf{A}_z &= (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{Z} \end{aligned} \tag{4}$$

where both $\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{S}^T\mathbf{S}$ are diagonal matrices with identical diagonal elements. Here, ICA readily solves the collinearity problem by guaranteeing an invertible matrix $\mathbf{S}^T\mathbf{S}$ because of the mutual orthonormality of the ICs.

Once $\mathbf{W}_z$ and the ICs are found, then the demixing matrix $\mathbf{W}_x$ and the mixing matrix $\mathbf{A}_x$ can be obtained from:

$$\begin{aligned} \mathbf{W}_x &= \mathbf{Q}^T\mathbf{W}_z \\ \mathbf{A}_x &= (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{X} = \mathbf{A}_z\mathbf{\Lambda}^{1/2}\mathbf{U}^T \end{aligned} \tag{5}$$

where $\mathbf{A}_x\mathbf{W}_x = \mathbf{I}$ and in the meantime it satisfies that $\mathbf{A}_x\mathbf{A}_x^T = \mathbf{A}_z\mathbf{\Lambda}\mathbf{A}_z^T$.

Based on the above formulation, for a set of spectra with J wavelengths acquired on N samples, $\mathbf{X}\ (J \times N)$, a common ICA model can be formulated as

$$\begin{aligned} \mathbf{S} &= \mathbf{X}\mathbf{W}_x \\ \mathbf{X} &= \mathbf{S}\mathbf{A}_x + \mathbf{E} \end{aligned} \tag{6}$$

where $\mathbf{S}\ (J \times R)$ is the estimated independent components from the observed spectra, which actually are the estimations of the source spectra in the mixture. The mixing matrix, $\mathbf{A}_x\ (R \times N)$, actually indicates the effects of the substances on mixture spectra. $\mathbf{E}\ (J \times N)$ is the unexplained residual.

2.2. Proposed Spectroscopic Modeling and Monitoring Strategy. Batch spectral measurement data are characterized by a three-way array. In each batch run, the spectral data can be arranged as a matrix with a size given by the number of wavelengths (J) $\times$ the number of time points (K). Then, spectral information collected from I similar batches can be organized as a three-way array $\underline{\mathbf{X}}\ (I \times J \times K)$. The final product information, for example, the concentrations of products, is collected in a two-way matrix $\mathbf{Y}\ (I \times J_y)$.

Here, it should be pointed out that one of the main challenges in the use of spectral measurements in comparison to engineering data is that spectroscopic data are generally more sensitive to changes in process conditions such as variations in the background medium, seasonal variations in feedstocks, or temperature fluctuations.[12] Spectroscopic effects such as light scattering and artifacts of the spectrometer such as baseline or wavelength shifts can also degrade quantitative accuracy. Wavelength selection has commonly been used as a means of focusing on regions of the spectrum immune to interferent variations. Mathematical preprocessing techniques, such as taking derivatives or orthogonal signal correction, can also be used to remove an undesired variation. In cases where spectral variation is due to known causes, such as instrumental laser drift or temperature influence, this can be corrected explicitly using direct standardization techniques.

As analyzed before, spectral measurements of batch processes are more likely to reflect the combination of chemical species present in the reaction mixture. Therefore, the measured spectra describe the chemical conditions relating to the molecular nature and the absolute concentrations of the species present within the system. The content of current reactants and their concentrations can reveal the underlying reaction kinetics. Over the course of time, different source species impose varying effects on the mixture spectra. For example, initially, the spectra are mainly determined by the pure reactants; after a reaction period, the spectra of the intermediates become visible and then slowly disappear again; and finally, the spectra are dominated by the end products. Alternation of the dominant species in the mixture and evolution of the chemical reaction are actually reflected externally by the changes of mixing coefficients and source components (i.e., ICs) estimated by ICA. The larger the magnitude of the mixing coefficients, the more important the estimated ICs. Therefore, in the present work, an ICA-based mixing coefficient matrix is used as the basic analysis unit from which the reference model is designed by revealing the batchwise normal stochastic fluctuations. Online monitoring is thus implemented to check whether the reaction is proceeding smoothly. Compared with the redundancy of wavelengths, the lower dimension of the mixing matrix provides a simpler and more convenient analysis platform.

2.2.1. Phase Identification. A multiplicity of phases is a common inherent feature of chemical batch processes. For multiphase batch processes, different phases are dominated by different chemical reaction mechanisms and phenomena, in which the underlying characteristics are similar within the same phase and dissimilar over different phases. The alternation of dominant source species and changes of underlying chemical reactions actually are synchronous with the evolution of the reaction phases. It will be beneficial to the comprehension of chemical reactions if different reaction phases can be identified and then a phase-based statistical modeling strategy can be developed to check different behaviors of various phases. Here, it should be noted that a "phase" in the context of the current work should be thought of as a "step" in the process, which is different from the conventional phase concept (gas, liquid, solid) in spectroscopy.

In the present work, the variant k-means clustering algorithm, which was designed by Lu et al.[30] and further developed by Zhao et al.,[31] is employed to classify the chemical reaction patterns. Here, the direct and basic clustering unit employed is the time-slice mixing matrix, which actually represents the underlying chemical characteristics and reveals the batchwise variability at each time. As mentioned before, the mixing relationship has a direct relationship to the concentration of constituent species, referring to reactants and intermediates, as well as products. Changes of mixing matrices, reflecting changes in the underlying chemical reaction mixture, can be used to determine the phases. The phase partition procedure is outlined in Figure 1a.

First, the three-way batch spectral data are organized as $\mathbf{X}\ (J \times IK)$, in which each column reveals the observed spectrum from each batch at each time.

Second, time-slice mean spectra are calculated as $\bar{\mathbf{X}}\ (J \times K)$ from $\mathbf{X}\ (J \times IK)$, to highlight the time-varying chemical information. This is feasible because these modeling batches are good and similarly driven by normal and acceptable variations over batches.

Third, ICA is run on the mean spectra $\bar{\mathbf{X}}\ (J \times K)$, to extract all of the underlying sources once they have appeared and reveal the time-varying underlying characteristics throughout the process duration from an overall viewpoint

$$\mathbf{S}(J \times R) = \bar{\mathbf{X}}\bar{\mathbf{W}} \tag{7}$$

where the sources, $\mathbf{S}\ (J \times R)$, summarize all spectra of species once they have emerged during the chemical reaction and the demixing coefficients, $\bar{\mathbf{W}}\ (K \times R)$, indicate their corresponding roles played in the mixture spectra along time direction.

Fourth, the time-slice mixing matrices, $\mathbf{A}_k\ (I \times R)$, are calculated directly from the time-slice spectral measurements,

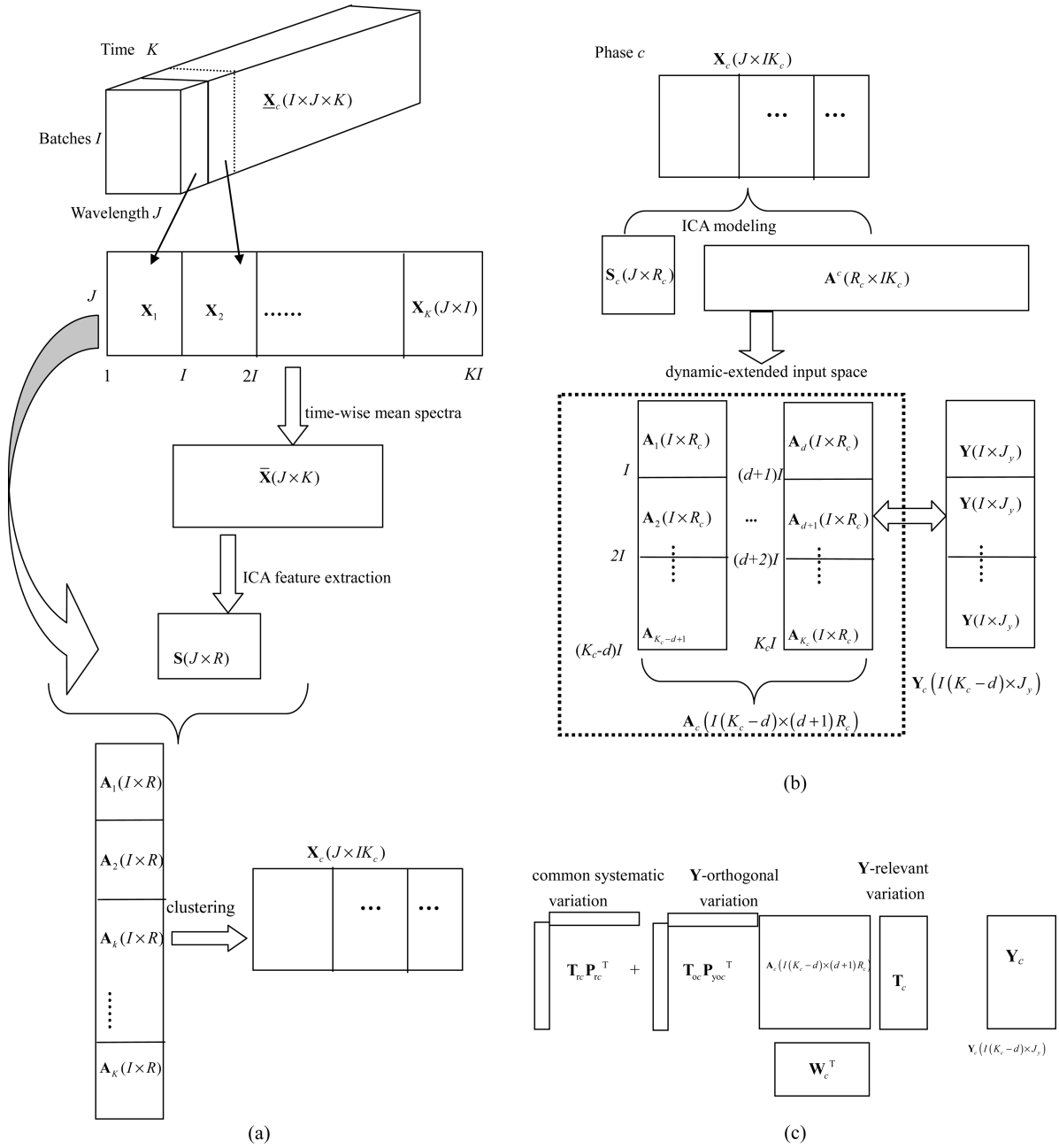

Figure 1. Scheme of the proposed method: (a) phase clustering, (b) phase-based data arrangement, (c) phase-based joint model structure.

$\mathbf{X}_k$ ($J \times I$), based on the estimated ICs, $\mathbf{S}$ ($J \times R$), and they are then normalized. This unveils the underlying characteristics at each time

$$\mathbf{A}_k(I\times R) = \left[(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{X}_k\right]^{\mathrm{T}} \tag{8}$$

Finally, C different phases are picked from the whole chemical reaction duration along the process time by clustering the normalized time slices $\mathbf{A}_k$ ($I \times R$), where each separated phase normally contains a series of successive samples.[30,31]

In this way, the consecutive reaction patterns with similar underlying chemical characteristics are collected together within the same phase, and dissimilar sample patterns are classified into different phases. Using a local model allows the underlying structures specific to each phase to be unveiled and thus benefits the detection of more specific locations of faults.

2.2.2. Phase-Based Data Arrangement. The following statistical analysis is performed focusing on each phase: The phase-specific spectral unit, $\mathbf{X}_c$ ($J \times IK_c$) (where K_c is the operation duration of phase c), is steadily obtained by placing all of the batchwise normalized time-slice spectral measurements within the same phase c, $\mathbf{X}^k$ ($J \times I$) ($k = 1, 2, ..., K_c$), next to each other, as shown in Figure 1b.

Subsequently, within each phase, the measurements are reanalyzed by ICA to extract the phase-representative sources, $\mathbf{S}_c$ ($J \times R_c$), and their mixing relationship, $\mathbf{A}^c$ ($R_c \times IK_c$)

$$\begin{aligned}\mathbf{S}_c &= \mathbf{X}_c\mathbf{W}^c\\ \mathbf{A}^c &= (\mathbf{S}_c{}^{\mathrm{T}}\mathbf{S}_c)^{-1}\mathbf{S}_c{}^{\mathrm{T}}\mathbf{X}_c\end{aligned} \tag{9}$$

In this way, the intermediates existing in the current phase are identified by $\mathbf{S}_c$, and the time-varying characteristics within the same phase are reflected by $\mathbf{A}^c$.

Timewise dynamics are common in chemical reaction processes in which the chemical reaction behaviors can correlate and influence each other along the time direction. To reveal such dynamics, the original mixing matrix is augmented to

Table 1. Modified CCA Modeling Algorithm

step	description
1	Calculate weight vectors $\mathbf{m}$ and $\mathbf{v}$ as the dominant eigenvectors of the matrices $[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{Y}[\mathbf{Y}^T\mathbf{Y}]^{-1}\mathbf{Y}^T\mathbf{X}$ and $[\mathbf{Y}^T\mathbf{Y}]^{-1}\mathbf{Y}^T\mathbf{X}[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{Y}$, respectively
2	Calculate the canonical variates $\mathbf{t} = \mathbf{Xm}$ and $\mathbf{u} = \mathbf{Yv}$ and scale them to unit length: $\mathbf{t} = \mathbf{t}/\|\mathbf{t}\|_2$ and $\mathbf{u} = \mathbf{u}/\|\mathbf{u}\|_2$
3	To normalize canonical variates, rescale the corresponding weight vectors: $\mathbf{w}$ and $\mathbf{v}$: $\mathbf{w} = \mathbf{w}/\|\mathbf{t}\|_2$ and $\mathbf{v} = \mathbf{v}/\|\mathbf{u}\|_2$
4	Calculate the corresponding loading vectors: $\mathbf{p}^T = \mathbf{t}^T\mathbf{X}$ and $\mathbf{q}^T = \mathbf{t}^T\mathbf{X}$
5	Deflate the input and output matrices: $\mathbf{X} = \mathbf{X} - \mathbf{tp}^T$ and $\mathbf{Y} = \mathbf{Y} - \mathbf{tq}^T$
6	Go to step 1 as another iteration with the deflated $\mathbf{X}$ and $\mathbf{Y}$ matrices for the next canonical variates
7	When the desired number of canonical vairates is obtained, output the weight matrices and loading matrices: $\mathbf{M}$ and $\mathbf{V}$, $\mathbf{P}$ and $\mathbf{Q}$
8	Calculate the CCA weight matrix to compute canonical variates $\mathbf{T}$ directly from the original $\mathbf{X}$: $\mathbf{R} = \mathbf{M}(\mathbf{P}^T\mathbf{M})^{-1}$

include time-lagged values and capture autocorrelations. As shown in Figure 1b, a new modeling descriptor region is defined as $\mathbf{A}_c$ $[I(K_c - d) \times (d + 1)R_c]$, where d is the chosen order of time lags. In fact, it can be also obtained as

$$\mathbf{A}_c{}^T = \begin{bmatrix} (\mathbf{S}_c{}^T\mathbf{S}_c)^{-1}\mathbf{S}_c{}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{S}_c{}^T\mathbf{S}_c)^{-1}\mathbf{S}_c{}^T & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (\mathbf{S}_c{}^T\mathbf{S}_c)^{-1}\mathbf{S}_c{}^T \end{bmatrix}_{(d+1)R_c \times (d+1)J} \cdot \begin{bmatrix} \mathbf{X}_{c-d} \\ \mathbf{X}_{c-d+1} \\ \vdots \\ \mathbf{X}_{c-0} \end{bmatrix} = \mathbf{G}\mathbf{X}_c{}^d \quad (10)$$

where the first term, $\mathbf{G}$, is a $(d + 1)R_c \times (d + 1)J$ matrix whose diagonal submatrices are the same and are all composed by $(\mathbf{S}_c{}^T\mathbf{S}_c)^{-1}\mathbf{S}_c{}^T$ and the second term, $\mathbf{X}_c{}^d$, is a d-order augmented spectral measurement matrix with dimension $(d + 1)J \times I(K_c - d)$, which is composed of $d + 1$ $\{J \times [I(K_c - d)]\}$-dimensional matrices. $\mathbf{X}_{c-d}$ covers those samples ranging from the beginning of the current phase to the end of the phase except for d sampling intervals; and $\mathbf{X}_{c-0}$ $\{J \times [I(K_c - d)]\}$ ranges from the $(d + 1)$th sample of the current phase to its end.

It is obvious that, based on the ICA mixing coefficients, the augmented model unit with certain time delays is much simpler and avoids the redundant problem resulting from lengthy wavelengths. Certainly, both within-batch and cross-batch dynamics can be readily revealed by the inclusion of time- and batchwise lags, as has been done in two-dimensional (2D) dynamic statistical analysis strategy.[32,33] In the present work, only timewise dynamics are considered and analyzed. Moreover, it should be pointed out that, unlike spectral measurements collected throughout each batch run, generally, the quality variable $\mathbf{Y}$ can be measured only at the end of each course. In this case, the numbers of rows in the quality variables are properly arranged to make a consistent sample size with descriptors. The I batches of normalized quality variables are duplicated by $I(K_c - d)$ within each phase, forming $\mathbf{Y}_c$ $[I(K_c - d) \times J_y]$.

Here, it should be pointed out that ICA is not the only choice for the recovery of pure spectra of constituent species and their concentrations from chemical mixture spectra. Another approach, called MCR as mentioned in the Introduction, can also be used as an alternative, which, however, is not addressed here.

2.2.3. Phase-Based Joint Modeling and Analysis Strategy. Because of the complexity of chemical reactions, there generally exist various types of variations under spectral measurements, and they have different influential effects on qualities. Not only the variable correlation in $\mathbf{X}$ space but also the correlations between $\mathbf{X}$ and qualities $\mathbf{Y}$[20] are important for chemical process monitoring. It would be more appealing if one could discriminate different process variations by relating $\mathbf{X}$ and $\mathbf{Y}$ and modeling them for online monitoring. This could provide comprehensive insight into the underlying characteristics. The key is to choose proper analysis algorithms to explore these variations. Neither single MPCA nor single MPLS can give a full description of the process variations. PCA[11,12] makes use of only the spectral information available in $\mathbf{X}$ space, which can tell the occurrence of process abnormality but not clarify whether it can influence the quality. PLS-based monitoring systems decompose $\mathbf{X}$ space under the supervision of $\mathbf{Y}$. However, it is also realized that, in the standard PLS system, the LVs include some systematic variations that are quality-irrelevant because the objective of PLS is to maximize the covariance information of the two spaces and a large covariance might not necessarily mean a strong correlation. That is, both $\mathbf{Y}$-related and $\mathbf{Y}$-irrelevant variations are mixed in the same monitoring statistic, T^2. Moreover, in the PLS residual part, some systematic variations are not explored.[34] To improve the PLS model, the orthogonal signal correction (OSC) idea[35−38] has been introduced. The concept behind all OSC algorithms is to work as a preprocessing step to remove systematic variations in $\mathbf{X}$ that are orthogonal to $\mathbf{Y}$. Trygg and Wold[38] put forward an orthogonal PLS (O-PLS) algorithm by directly integrating OSC into the regular PLS algorithm, which separates the quality-orthogonal systematic variation from the original PLS LVs. Further O2-PLS method[39] was presented by the same authors, which carried out OSC filtering focusing on both the response space $\mathbf{Y}$ and the input space $\mathbf{X}$. Alternatively, Yu and MacGregor[40] employed canonical correlation analysis (CCA)[41,42] as a postprocessing tool to directly exploit and maximize the correlation between the PLS LV subspace and the $\mathbf{Y}$ space. It can overcome the general rank-

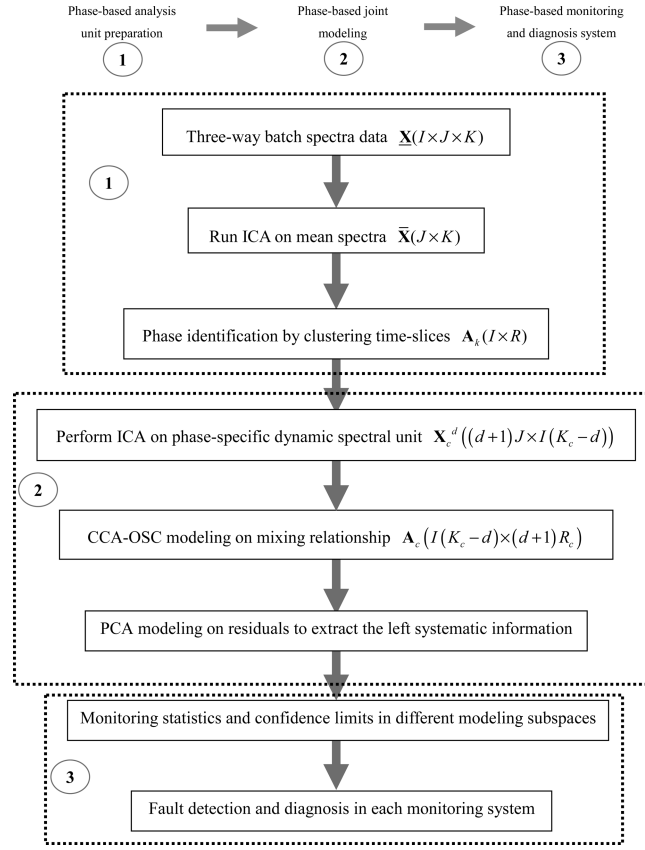

Figure 2. Flow diagram of the proposed modeling strategy.

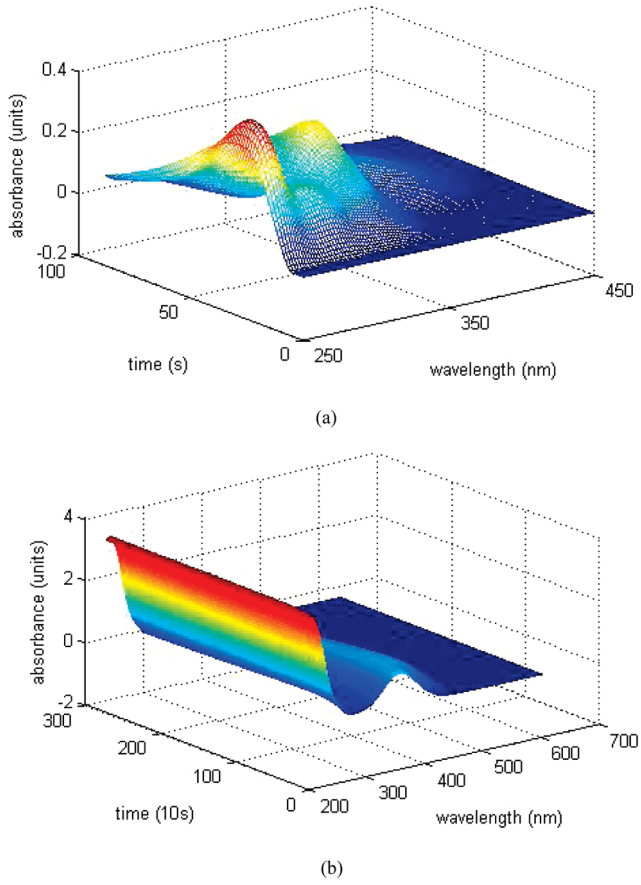

Figure 3. Profile of mean spectra in (a) case study 1 and (b) case study 2.

deficient problem existing in the single CCA algorithm resulting from the involved calculation $(\mathbf{X}^T\mathbf{X})^{-1}$. However, the above algorithms are all employed to design and improve the calibration model for quality prediction. Their role in deriving different systematic variations for process monitoring has not been explored.

On the basis of the above analysis, in this work, CCA, OSC, and PCA algorithms are combined to design a joint monitoring system. Its purpose is to acquire a further decomposition of spectral $\mathbf{X}$ space into four subspaces, in which different types of variations are explored for monitoring. The first part is exploited by CCA, revealing the process variations closely related to quality. Here, it should be pointed out that, in the current work, the high-dimensional wavelength of the original spectral measurement has been replaced by an R-dimensional mixing relationship, $\mathbf{A}_c$ $[I(K_c - d) \times (d + 1)R_c]$, which thus avoids the rank-deficient problem. This provides a feasible analysis platform for the direct use of the CCA algorithm to reveal the correlation information between two spaces. The second one is explained by OSC, covering the real quality-orthogonal systematic variation. Further, the original residuals after analysis of CCA−OSC are split into two different parts by PCA: common systematic variations and the final residual. Then, different monitoring statistics are developed, and fault detection is thus performed in different parts of $\mathbf{X}$ space. They jointly work together to reveal the abnormal behaviors imposed on different variations, which will make the monitoring results more convincing and meaningful. The analysis and modeling scheme is shown in Figure 1c.

First, the original CCA algorithm is modified[43] so that proper statistical indices can be designed for the specific purpose of monitoring, as shown in Table 1. It is used to extract the most quality-related systematic variation information within each phase

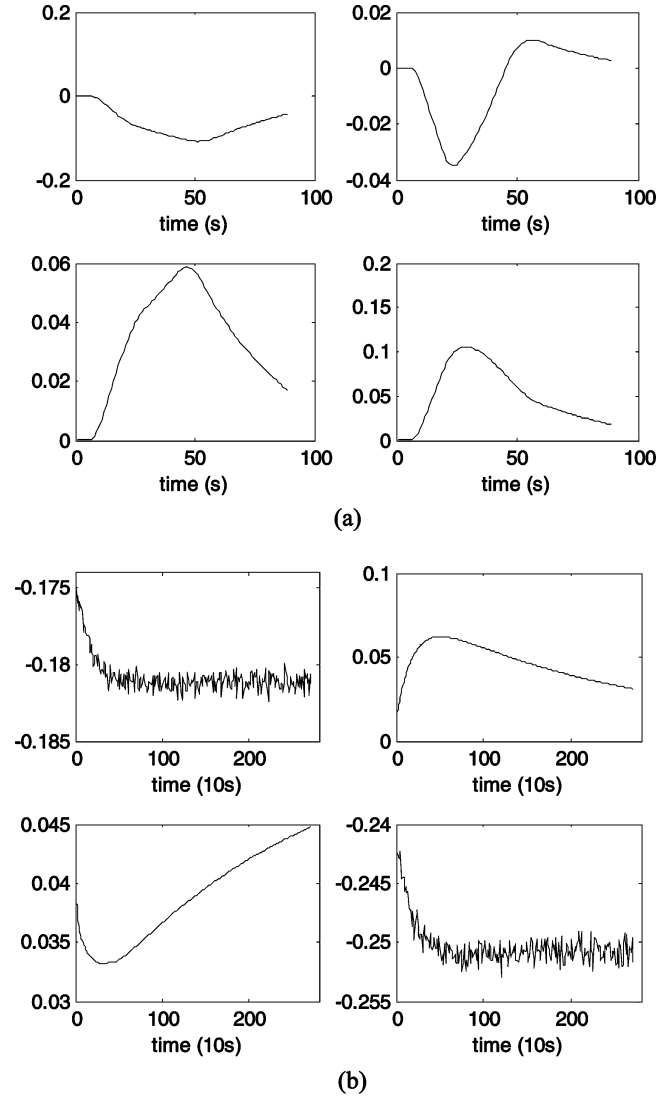

Figure 4. Time-varying trajectory of ICA mixing relationship estimated from mean spectra in (a) case study 1 and (b) case study 2.

$$\begin{aligned} \mathbf{T}_{cc} &= \mathbf{A}_c\mathbf{W}_{cc} \\ \mathbf{A}_c &= \mathbf{T}_{cc}\mathbf{P}_{cc}^{\ T} + \mathbf{E}_{1c} \end{aligned} \tag{11}$$

where $\mathbf{T}_{cc}$ $[I(K_c - d) \times A_c]$ is the scores matrix of systematic variations covarying between the phase-representative descriptor space $\mathbf{A}_c$ and the process-overall $\mathbf{Y}$, $\mathbf{W}_{cc}$ $[(d + 1)R_c \times A_c]$ is the weights matrix, and $\mathbf{P}_{cc}$ $[(d + 1)R_c \times A_c]$ is the loadings matrix of $\mathbf{A}_c$. $\mathbf{E}_{1c}$ is the initial residual after first-step CCA.

Second, the OSC algorithm[36] is performed on the CCA residual to extract the mutually irrelevant systematic information in both input and output spaces

$$\begin{aligned} \mathbf{T}_{oc} &= \mathbf{E}_{1c}\mathbf{W}_{yoc} \\ \mathbf{E}_{1c} &= \mathbf{T}_{oc}\mathbf{P}_{yoc}^{\ T} + \mathbf{E}_{2c} \end{aligned} \tag{12}$$

where $\mathbf{T}_{oc}$ $[I(K_c - d) \times A_{oc}]$ is the scores matrix of orthogonal systematic variations in $\mathbf{A}_c$, $\mathbf{W}_{yoc}$ $[(d + 1)R_c \times A_{oc}]$ represents the weights of orthogonal parts, and $\mathbf{P}_{yoc}$ $[(d + 1)R_c \times A_{oc}]$ represents their loadings. $\mathbf{E}_{2c}$ $[I(K_c - d) \times (d + 1)R_c]$ is the residual after CCA−OSC model interpretation. Here, OSC works not as a pretreatment prior to PLS modeling, but as an

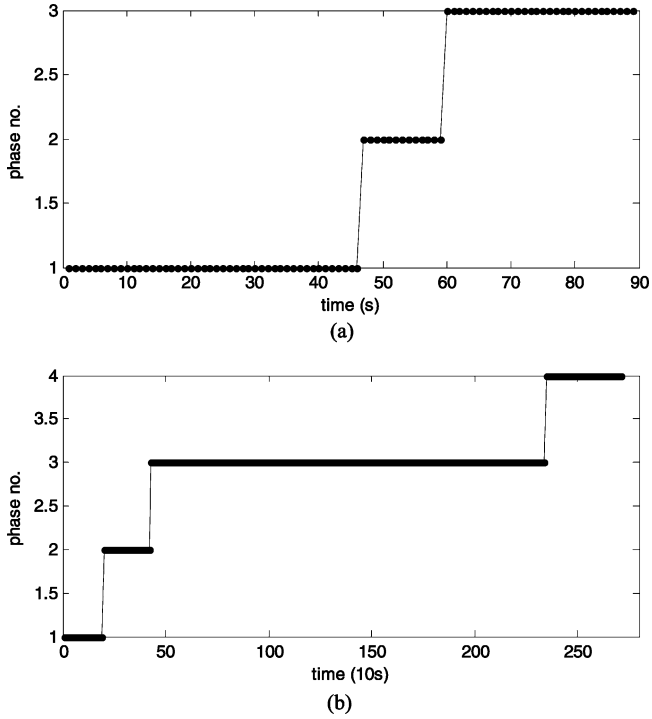

Figure 5. Phase identification result in (1) case study 1 and (b) case study 2.

individual feature extraction approach to reveal certain types of systematic information existing in spectral data, and thus, it forms one variation subspace for process monitoring.

After CCA−OSC modeling, systematic information is still left in $\mathbf{E}_{2c}$ and can be further extracted by PCA

$$\begin{aligned}\mathbf{T}_{rc} &= \mathbf{E}_{2c}\mathbf{P}_{rc}\\ \mathbf{E}_{2c} &= \mathbf{T}_{rc}\mathbf{P}_{rc}{}^{T} + \mathbf{E}_{c}\end{aligned} \tag{13}$$

where $\mathbf{T}_{rc}$ $[I(K_c - d) \times A_{rc}]$ is the PCA scores matrix and $\mathbf{P}_{rc}$ $[(d + 1)R_c \times A_{rc}]$ is the corresponding PCA loadings matrix. $\mathbf{E}_c$ is the PCA residual, that is, the final residual, which thus reflects only the random noises.

In conclusion, using the above joint modeling algorithm, we can finally model the chemical reaction information within each phase as follows

$$\mathbf{A}_c = \mathbf{T}_c\mathbf{P}_c{}^{T} + \mathbf{T}_{oc}\mathbf{P}_{yoc}{}^{T} + \mathbf{T}_{rc}\mathbf{P}_{rc}{}^{T} + \mathbf{E}_c \tag{14}$$

Compared with the LV single model, the proposed phase-specific LV joint model structure, which integrates the CCA, OSC, and PCA algorithms, is clearer and more comprehensive for describing the variations in $\mathbf{X}$ space.

Here, comments should be in reference to the number of retained ICs. As described above, the proposed method actually uses a two-step feature extraction procedure: the first-step ICA and the second-step joint modeling. In the first-step modeling, ICA prepares the analysis unit, that is, the mixing relationship. In the second-step modeling, the joint modeling algorithm (CCA−OSC−PCA) focuses on the mixing coefficients to derive the important information by feature extraction. Generally, if a smaller number of ICs is kept in the first step, some underlying information might be overlooked, and thus, the designed monitoring system might not comprehensively describe the operation status. In contrast, if more ICs are retained, some of them might be more likely to be noises, which, however, does not necessarily impose a significant influence on the modeling performance. If the ICs play a considerable role in the spectral measurement, the associated mixing relationships also cover information on large systematic variations. In contrast, for the unimportant ICs that are more likely to stand for only noises, their mixing relationships neither follow certain rules nor cover sufficient information. The second-step joint modeling can check the roles of their associated mixing coefficients. The noise information is excluded, and only the important features are retained. In the current work, the number of retained ICs can be determined by trial and error to obtain a least false alarm rate for normal cases.

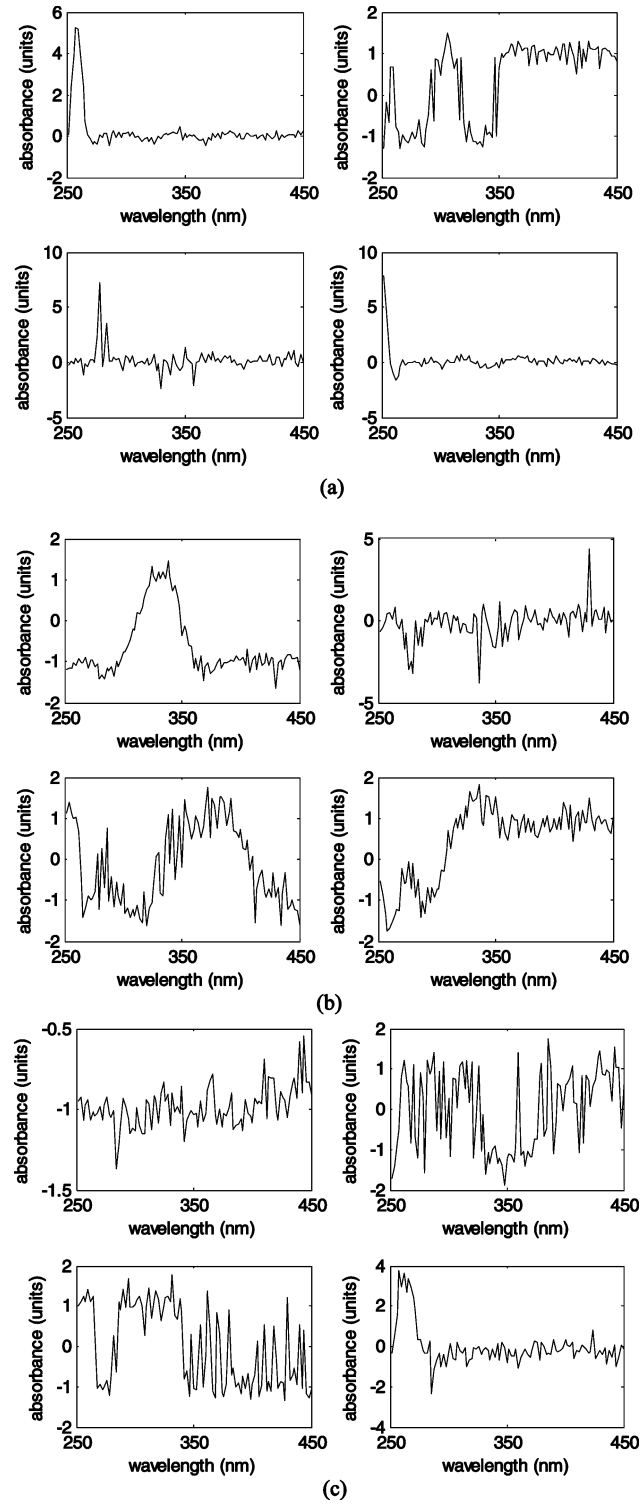

Figure 6. Spectral profile of the first four ICs extracted in each phase in case study 1: (a) phase I, (b) phase II, (c) phase III.

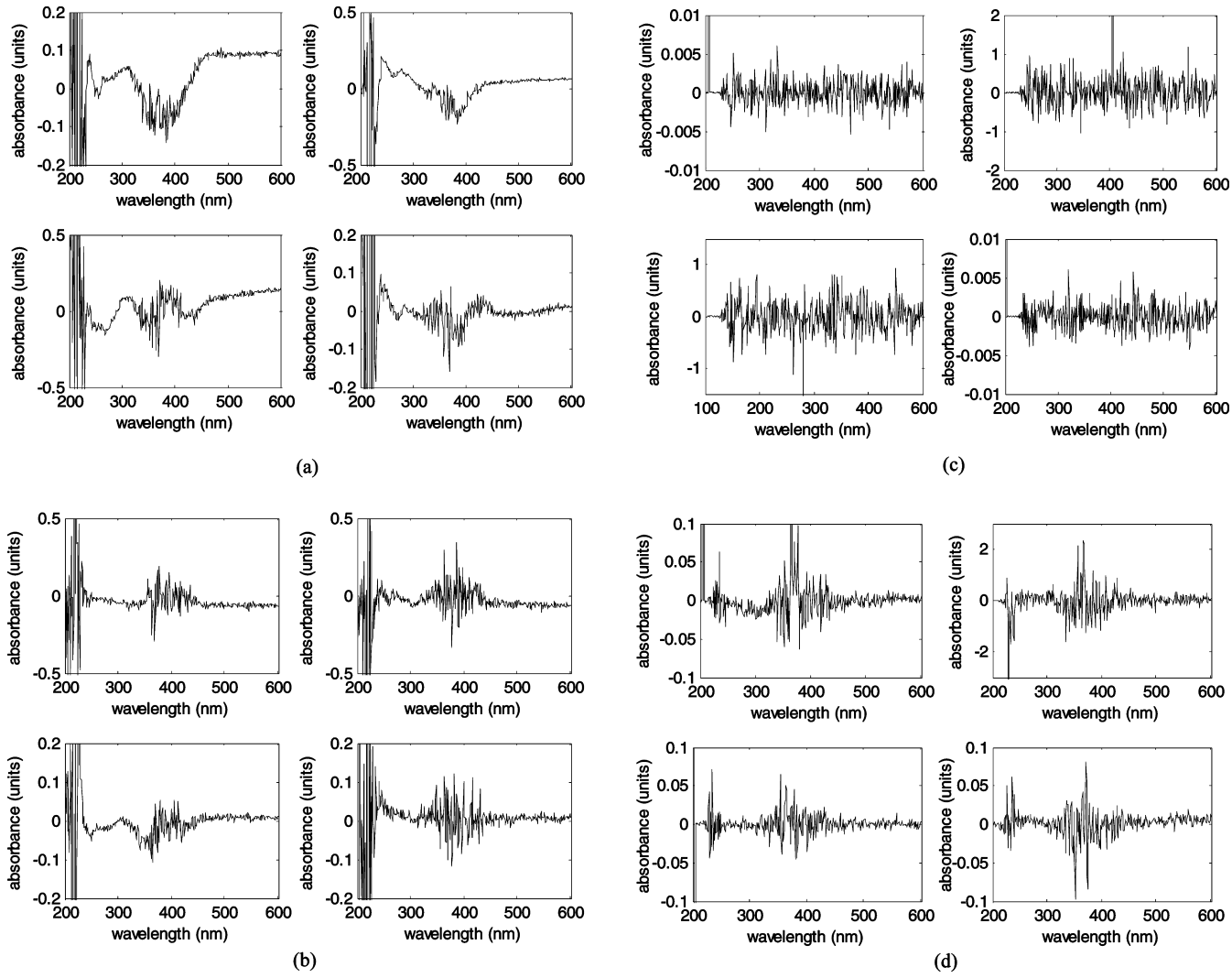

Figure 7. Spectral profile of the first four ICs extracted in each phase in case study 2: (a) phase I, (b) phase II, (c) phase III, (d) phase IV.

2.2.4. Monitoring Statistics and Confidence Limits. For monitoring purposes, two types of statistics are commonly calculated to exploit the measured $\mathbf{X}$: the T^2 statistic, which describes the systematic part captured by monitoring models, and the Q statistic, which represents the residual part unoccupied by monitoring models. In the joint structures, $\mathbf{T}_{cc}$, $\mathbf{T}_{oc}$, and $\mathbf{T}_{rc}$ contain information on the systematic variations of chemical reactions located in the first three subspaces and thus are employed for the T^2 statistic, whereas the residual, $\mathbf{E}_c$, is suitable for the SPE statistic, the squared prediction error,

$$\begin{aligned} T_{ci,k}^{\ 2} &= (\mathbf{t}_{ci,k} - \bar{\mathbf{t}}_{ck})^{\mathrm{T}}\mathbf{O}_{cc}^{\ -1}(\mathbf{t}_{ci,k} - \bar{\mathbf{t}}_{ck}) = \mathbf{t}_{ci,k}^{\ \mathrm{T}}\mathbf{t}_{ci,k} \\ T_{oi,k}^{\ 2} &= (\mathbf{t}_{oi,k} - \bar{\mathbf{t}}_{ok})^{\mathrm{T}}\mathbf{O}_{oc}^{\ -1}(\mathbf{t}_{oi,k} - \bar{\mathbf{t}}_{ok}) = \mathbf{t}_{oi,k}^{\ \mathrm{T}}\mathbf{O}_{oc}^{\ -1}\mathbf{t}_{oi,k} \\ T_{ri,k}^{\ 2} &= (\mathbf{t}_{ri,k} - \bar{\mathbf{t}}_{rk})^{\mathrm{T}}\mathbf{O}_{rc}^{\ -1}(\mathbf{t}_{ri,k} - \bar{\mathbf{t}}_{rk}) = \mathbf{t}_{ri,k}^{\ \mathrm{T}}\mathbf{O}_{rc}^{\ -1}\mathbf{t}_{ri,k} \\ \mathrm{SPE}_{i,k} &= e_{i,k}^{\ \mathrm{T}}e_{i,k} \end{aligned} \tag{15}$$

where $\mathbf{t}_{ci,k}$ ($A_c \times 1$), $\mathbf{t}_{oi,k}$ ($A_{oc} \times 1$), $\mathbf{t}_{ri,k}$ ($A_{rc} \times 1$), and $e_{i,k}$ are the canonical variate, the OSC component, the principal component, and the residual vector, respectively, at the kth time for the ith batch; $\bar{\mathbf{t}}_{ck}$, $\bar{\mathbf{t}}_{ok}$, and $\bar{\mathbf{t}}_{rk}$ denote the corresponding mean vectors, which are all zero vectors because of the use of time-slice mean-centering in the data preprocessing procedure; and $\mathbf{O}_{cc}$, $\mathbf{O}_{oc}$, and $\mathbf{O}_{rc}$ are the phase-based variance-covariance matrices of components, respectively, in which $\mathbf{O}_{cc}$ is actually an identity matrix resulting from the unit variance requirement of canonical vectors.

The premise of a Gaussian distribution provides an important basis for deriving the confidence limits of monitoring statistics. Therefore, the phase-representative T^2 control limits in each system subspace can be defined by the F distribution with an α significance factor.[2,44] In the error subspace, the representative confidence limit of the SPE within each phase can be approximated by a weighted chi-squared distribution,[2] $g\chi_h^2$.

When a new batch spectral sample is available within the cth phase as indicated by time, a d-order augmented analysis unit, $\mathbf{a}_{\mathrm{new}}$ [$(d + 1)R_c \times 1$] is calculated as

$$\mathbf{a}^{\mathrm{new}} = \mathbf{G}\begin{bmatrix} \mathbf{x}_{\mathrm{new}-d} \\ \mathbf{x}_{\mathrm{new}-d+1} \\ \vdots \\ \mathbf{x}_{\mathrm{new}} \end{bmatrix} = \mathbf{G}\mathbf{x}_{\mathrm{new}}^{\ d} \tag{16}$$

where

$$\mathbf{x}_{\mathrm{new}}^{\ d} = \begin{bmatrix} \mathbf{x}_{\mathrm{new}-d} \\ \mathbf{x}_{\mathrm{new}-d+1} \\ \vdots \\ \mathbf{x}_{\mathrm{new}} \end{bmatrix}$$

is a d-order [$(d + 1)J \times 1$]-dimensional augmented spectral measurement vector by including previous ones prior to the current time within the same batch.

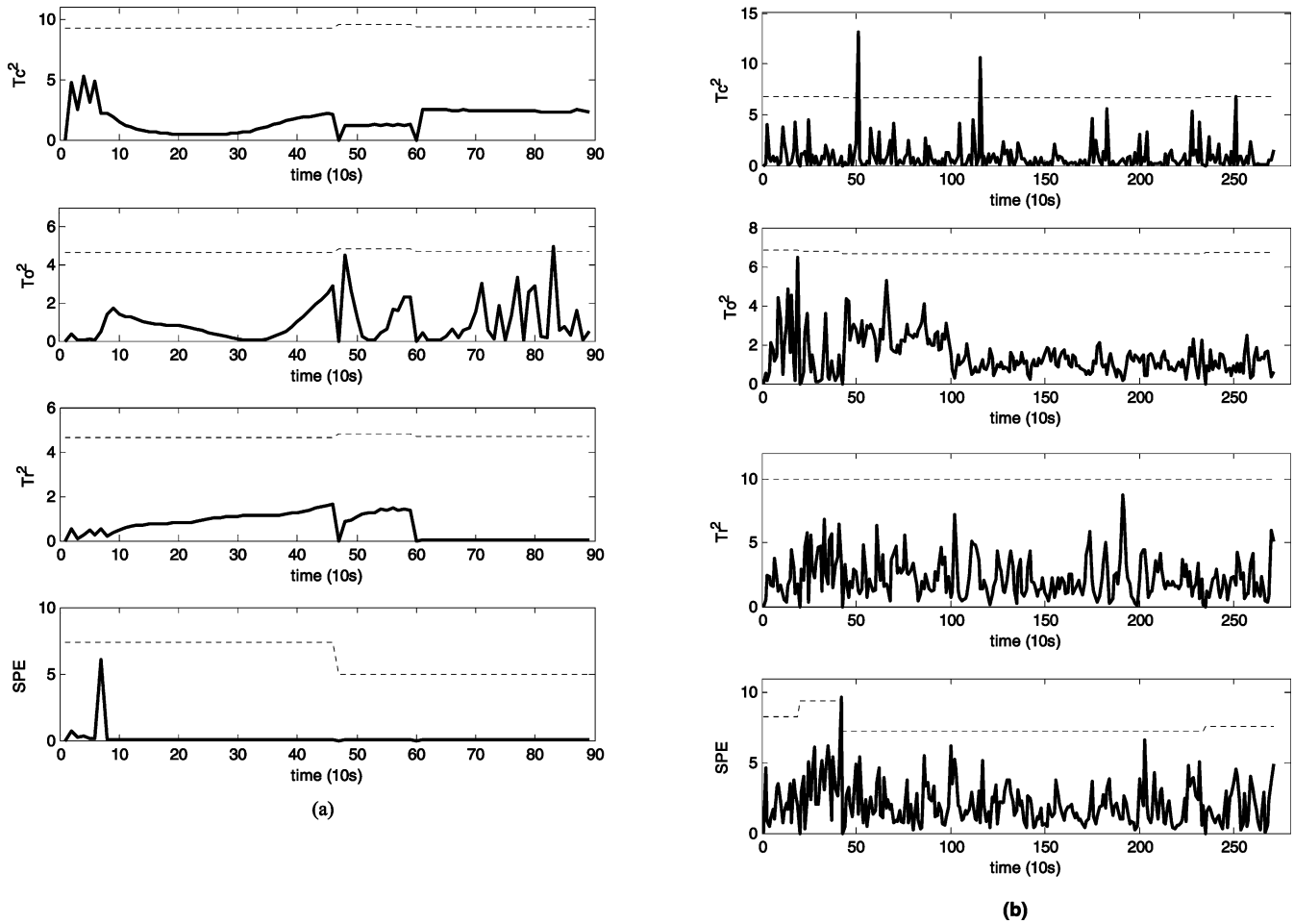

Figure 8. Monitoring result for one normal batch in (a) case study 1 and (b) case study 2 (dashed line, 99% control limit; solid line, online monitoring statistics).

Subsequently, it is projected onto the joint model structures into which the current time falls

$$\begin{aligned}
\mathbf{t}_{\text{cnew}}{}^{\text{T}} &= \mathbf{a}_{\text{new}}{}^{\text{T}}\mathbf{W}_{cc} \\
\mathbf{e}_{1\text{new}} &= \mathbf{a}_{\text{new}} - \mathbf{P}_{cc}\mathbf{t}_{\text{cnew}} \\
\mathbf{t}_{\text{onew}}{}^{\text{T}} &= \mathbf{e}_{1\text{new}}{}^{\text{T}}\mathbf{W}_{yoc} \\
\mathbf{e}_{2\text{new}} &= \mathbf{e}_{1\text{new}} - \mathbf{P}_{yoc}\mathbf{t}_{\text{onew}} \\
\mathbf{t}_{\text{rnew}}{}^{\text{T}} &= \mathbf{e}_{2\text{new}}{}^{\text{T}}\mathbf{P}_{rc} \\
\mathbf{e}_{\text{new}} &= \mathbf{e}_{2\text{new}} - \mathbf{P}_{rc}\mathbf{t}_{\text{rnew}}
\end{aligned} \tag{17}$$

The T^2 and SPE statistics are thus calculated as

$$\begin{aligned}
T_{\text{cnew}}{}^2 &= \mathbf{t}_{\text{cnew}}{}^{\text{T}}\mathbf{t}_{\text{cnew}} \\
T_{\text{onew}}{}^2 &= \mathbf{t}_{\text{onew}}{}^{\text{T}}\mathbf{O}_{ok}{}^{-1}\mathbf{t}_{\text{onew}} \\
T_{\text{rnew}}{}^2 &= \mathbf{t}_{\text{rnew}}{}^{\text{T}}\mathbf{O}_{rk}{}^{-1}\mathbf{t}_{\text{rnew}} \\
\text{SPE}_{\text{new}} &= \mathbf{e}_{\text{new}}{}^{\text{T}}\mathbf{e}_{\text{new}}
\end{aligned} \tag{18}$$

Process monitoring is conducted by continuously comparing all four statistics with the predetermined control limits. If they stay well within the predefined normal regions, the chemical reaction can be deemed to be normal. Otherwise the statistics will go beyond the control limits, indicating the occurrence of abnormal variations. The four monitoring charts in response to different types of variations according to $\mathbf{Y}$ can jointly reveal more abundant variation information.

2.2.5. Fault Diagnosis. The contribution values of mixing coefficients to different monitoring indices are calculated here to isolate the unusual variations within each phase, which are simpler to calculate and easier to read than those based on redundant wavelengths. Considering that the mixing coefficients are directly associated with the estimated ICs, corresponding to the abnormal mixing coefficients, it is reliable to directly point out the concerned ICs. Then, by examining the wavelength profile of the concerned ICs, it is convenient to investigate the effects of abnormal variations on the mixture spectra. Further, if one can know beforehand about the pure spectra of analytes and directly relate these ICs to the real constituent species existing in the chemical mixture, it is more direct and clear to determine what is really wrong with the chemical reaction. Because of the imposed independence of the estimated spectra of constituent species, the identification results should be more convincible. This is one advantage compared with MPCA/MPLS, in which the correlations among spectral wavelengths can more or less influence each other, and thus, the identified spectral region that is most influenced by the abnormality might not accurate. Moreover, because of the separation of different variations, once a fault is detected in some subspace, the type of fault variation can be located specifically. An experienced operator can accordingly assign the possible cause more easily. According to the above-mentioned analyses, the proposed method provides a potential for more accurate fault diagnosis.

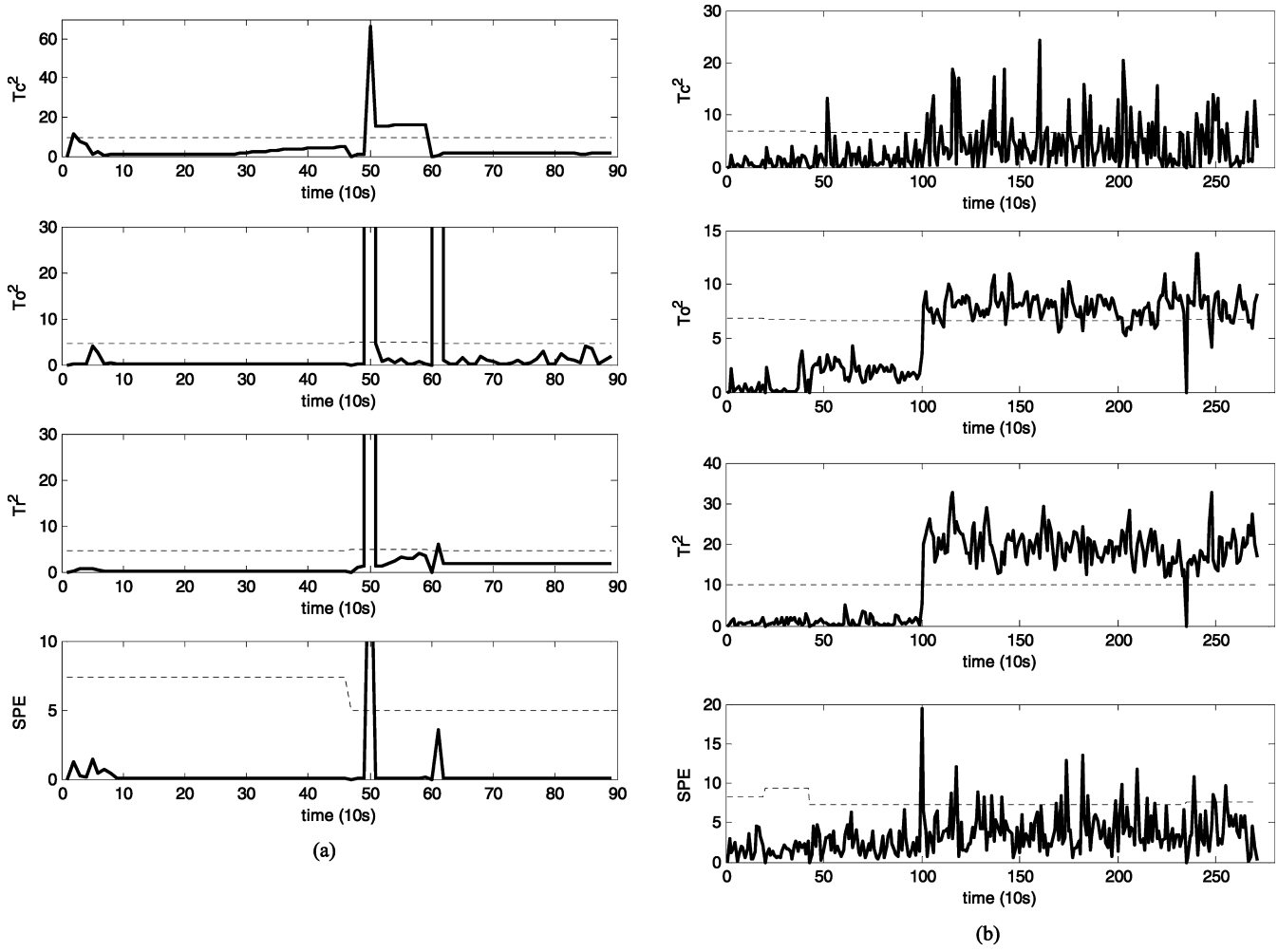

Figure 9. Monitoring results for the abnormal batch in (a) case study 1 and (b) case study 2 (dashed line, 99% control limit; solid line, online monitoring statistics).

The contribution to the T^2 statistic can be defined by investigating the individual effect of each single mixing coefficient on the statistics

$$\begin{aligned} C_{T_c^2,r} &= \mathbf{t}_{\text{cnew}}{}^{\text{T}}\mathbf{w}_{cc,r}a_{\text{new},r} \\ C_{T_o^2,r} &= \mathbf{t}_{\text{onew}}{}^{\text{T}}\mathbf{O}_{ok}{}^{-1}\mathbf{w}_{\text{yoc},j}e_{1\text{new},r} \qquad r = 1, 2, ...,(d+1)R_c \\ C_{T_r^2,r} &= \mathbf{t}_{\text{rnew}}{}^{\text{T}}\mathbf{O}_{rk}{}^{-1}\mathbf{p}_{rc,r}e_{2\text{new},r} \end{aligned} \tag{19}$$

where the vector $\mathbf{w}_{cc,r}$ comes from the rth row of the weights matrix $\mathbf{W}_{cc}$ and is similar to $\mathbf{w}_{\text{yoc},r}$ and $\mathbf{p}_{rc,r}$, which all correspond to the concerned rth mixing coefficient. $a_{\text{new},r}$ is the rth element of the new mixing vector $\mathbf{a}_{\text{new}}$ $[(d+1)R_c \times 1]$, and it is similar to $e_{1\text{new},r}$ and $e_{2\text{new},r}$.

Moreover, the contribution to SPE is also captured based on the same fact that the contributions summed over all mixing coefficients equal the new SPE value

$$C_{\text{spe},r} = e_{\text{new},r}{}^2 \tag{20}$$

where $e_{\text{new},r}$ is the final residual corresponding to the rth mixing value.

A simplified flow diagram describing the proposed modeling procedure is shown in Figure 2 to illustrate how it is implemented.

3. Illustration and Discussion

3.1. Experimental Data Sets. Case Study 1: Flow Injection Analysis (FIA) System. This spectral data set considers three different analytes in a flow injection analysis (FIA) system.[45] The data are available online at http://newton.foodsci.kvl.dk/research/data/Flow_Injection. The absorbance was detected by a diode-array detector from 250 to 450 nm in 2-nm intervals. The absorption spectrum was determined each second 89 times during one injection. Twelve runs of injections were implemented. Therefore, the spectral data set was organized into a 12 (batches) × 100 (wavelengths) × 89 (times) array. The three analytes present in the sample were 2-, 3-, and 4-hydroxybenzaldehyde (HBA). The three analytes have different absorption spectra. The concentrations of the three analytes in the 12 batches were used as the quality matrix $\mathbf{Y}$ (12 × 3).

Case Study 2: Two-Step Consecutive Chemical Reaction. This data set comes from a two-step consecutive reaction that has been described previously in the literature.[46] During this reaction, 3-chlorophenylhydrazonopropane dinitrile (A), an uncoupler of oxidative phosphorylation in cells, and 2-mercaptoethanol (B) form an intermediate adduct (C). Then, the adduct is hydrolyzed in an apparently intramolecular reaction to give the main product 3-chlorophenylhydrazono-

cyanoacetamide (D) and the byproduct ethylene sulfide (E). The reaction scheme is

$$A + B \xrightarrow{k_1} C \xrightarrow{k_2} D + E \quad (21)$$

A Hewlett-Packard 8453 UV−vis spectrophotometer with a diode-array detector was used to monitor the reaction. The UV−vis spectra for each run were recorded evenly every 10 s, and 271 samples were obtained. They revealed different reaction durations from batch to batch because they were measured from different starting times. The wavelength range used was 200−600 nm (wavelength resolution is 1 nm). Therefore, the spectroscopic data of each batch were arranged as a matrix with a size of number of wavelengths (401) × number of time points (271). Ten successful batch runs were performed, which were used as training data for developing the monitoring system. The three-way spectral measurement was recorded as $\mathbf{X}$ ($I \times J \times K$), where I is the number of batches, J is the number of wavelengths, and K is the number of time intervals. The actual reaction time showed a normal batchwise variation and was used as the response variable $\mathbf{Y}$ ($I \times 1$). Therefore, by relating $\mathbf{X}$ and $\mathbf{Y}$, we can obtain the changes in process status evolving along the reaction progress. The measurements are available online at http://www.bdagroup.nl/downloads/bda_downloads.html.

3.2. Simulation Methodology and Results. First, Figure 3 shows the profile of mean spectra over the modeling batches throughout the reaction duration for both case studies. ICA was performed on the mean spectra, and the time-varying trajectories of mixing coefficients corresponding to the first four ICs are shown in Figure 4, which indicate the changes in concentrations of the estimated constituents in the mixtures. By clustering, phases are thus identified from the chemical reaction process, as shown in Figure 5. The results reveal that, without any prior process knowledge, both chemical processes can be automatically divided into several phases. For the first case, three phases were obtained, and for the latter case, four were indicated, with longer and shorter durations. Without losing generality, each phase suggests different underlying chemical reaction information, as well as different mixing relationships.

Then, focusing on each phase, the dominant ICs were re-extracted. As an example, the spectral profiles of the first four ICs in different phases are shown in Figures 6 and 7, respectively, for the two cases. Possible noise variations are revealed during some wavelength regions, which can be eliminated by wavelength selection. The linear combination of ICs constructs the mixture spectral measurement, which is useful for fault identification. If the disturbance-subject ICs are identified, their effects on the mixture spectra can be readily revealed according to their profiles and the associated mixing relationships. Setting the time delay $d = 1$, phase-based joint LV-based statistical models are thus designed for process analysis and online monitoring. Cross-validation recommends that different latent components are needed for each phase to describe different systematic variations. During CCA modeling, for the first case study, two canonical components are related to multivariable responses in each phase; for the second case, only one canonical component is retained within each phase, which is most correlated with the single-variable response.

Under the same operating conditions as the reference batches, we generated both normal and abnormal batches as testing batches that can be used to verify the monitoring performance of the designed joint statistical models. For the normal case, batchwise normally distributed noise signals with zero mean were imposed. The noises accounted for the size of normal stochastic variability at each wavelength over batches, which were regarded to be acceptable. Figure 8 displays the monitoring results of normal batches in the two case studies. Generally, all of the monitoring statistics stay well below the control limits, indicating that different types of variations all progress normally and the current batch coincides with the historical benchmark of normal operation. To generate fault batches, random disturbances with twice the normal batchwise variability were imposed on the spectral measurement. Abnormality spans the 50−60 time region and the 1−100-nm wavelength region in the first case study, and it starts from the 100th sampling time (phase III) onward and spans the 200−400-nm wavelength region in the second case study. This makes the current batch behave beyond the expected common-cause variation trajectory captured in the NOC (normal operating condition) regions. Monitoring results are shown in Figure 9. For the first case study, clear and stable alarms are revealed, especially by the T_c^2 monitoring system, which means that the abnormal behavior mainly disturbs the quality-related systematic information. Comparatively, if variations were not separated, one would not be sure whether the fault affected product quality even if the fault could be detected. For the second case study, all types of systematic information were more or less influenced by undesirable disturbances. T_o^2

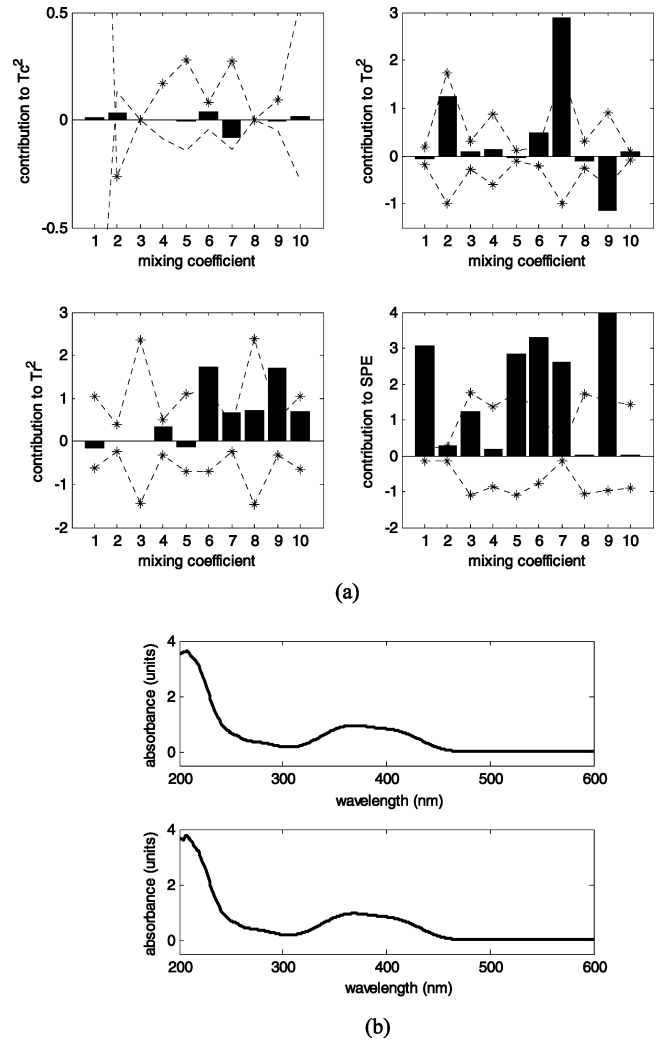

Figure 10. (a) Fault diagnosis result in case study 2 (− ∗ −, control limit of contribution values). (b) Spectra comparison between the normal and fault cases.

Table 2. Quantitative Description of Variations in X Space

model statistics	$R^2\mathbf{X}_{cc}$ (%)	$R^2\mathbf{X}_{oc}$ (%)	$R^2\mathbf{X}_{rc}$ (%)	$R^2\mathbf{E}_c$ (%)
calculation	$[\sum(\mathbf{T}_c\mathbf{P}_c^T)^2]/(\sum\mathbf{A}_c^2)$	$[\sum(\mathbf{T}_{oc}\mathbf{P}_{yoc}^T)^2]/(\sum\mathbf{A}_c^2)$	$[\sum(\mathbf{T}_{rc}\mathbf{P}_{rc}^T)^2]/(\sum\mathbf{A}_c^2)$	$[\sum(\mathbf{E}_c)^2]/(\sum\mathbf{A}_c^2)$
phase				
I	16.87	18.83	37.70	26.61
II	11.33	18.89	40.15	29.63
III	5.85	25.08	45.66	23.42
IV	7.42	24.11	44.29	24.18

and T_r^2 yield more obvious out-of-control behaviors, which indicates that the current disturbance are more influential on the two types of systematic variations. When alarms are generated, the contribution plot defined in subsection 2.2.5 is employed to interrogate the assignable cause. Taking an example for the second case study, a possible fault cause was synthetically checked at the 100th sampling time, especially focusing on the T_o^2 and T_r^2 statistical indices. From the contribution plot shown in Figure 10a, it can be seen that the sixth, seventh, and ninth sampling times in the *d*-order time-lags mixing vector (i.e., the first, second, and fourth ICs at the 100th sampling interval) are responsible for the abruptly changing statistics. Referring to the wavelength profiles of the concerned ICs shown in Figure 7c, one can determine the effects of operation faults on the mixture spectra. Actually, their flat spectra generally agree well with the real situation that the difference is small in comparing the mixture spectra of the normal case with those of the fault case, as shown in Figure 10b. Potentially, according to the spectra of the responsible ICs and the types of disturbed systematic variations, experienced operators can quickly determine the fault location. Especially if the faulty ICs can be related to the real constituents, this will provide more information helpful to fault isolation.

Moreover, corresponding to the four different subspaces, some model statistics can be defined here to obtain a further quantitative description of different **X** variations. Here, it is remarkable that they are not directly used for process monitoring but are helpful to further process analysis and algorithm understanding. They include quality-relevant variations by CCA, $R^2\mathbf{X}_{cc}$; quality-orthogonal variations by OSC, $R^2\mathbf{X}_{oc}$; common systematic variations by PCA, $R^2\mathbf{X}_{rc}$; and the final residuals, $R^2\mathbf{E}_c$. Their calculations are shown in Table 2, taking an example for the second case. From the results, in addition to the most quality-related information ($R^2\mathbf{X}_{cc}$) and the strictly quality-orthogonal information ($R^2\mathbf{X}_{oc}$), the remaining descriptors are still informative with a certain amount of variation ($R^2\mathbf{X}_{rc}$). Therefore, the third-step PCA is necessary to account for the non-negligible amount in each phase. From the above analyses, it is well realized that different types of chemical reaction variations exist with different amounts. The separation is useful and meaningful for the deep comprehension of underlying chemical reaction information. A joint modeling strategy is thus necessary to comprehensively monitor the reaction status.

From the preceding discussion, one can see that a series of generally favorable statistical analysis and encouraging monitoring results have demonstrated the effectiveness of the proposed method for spectroscopic analysis to improve process understanding and online monitoring. Its superiority is especially obvious for multiphase chemical reaction processes with different underlying variation characteristics over different phases, as well as timewise dynamics within each phase.

4. Conclusions

In this article, an improved modeling and monitoring strategy is presented for spectroscopy analysis. Rather than directly focusing on redundant wavelengths, this approach first extracts the underlying independent components and their mixing relationships by ICA, which simplifies the subsequent calculation complexity and enhances model interpretation. Accordingly, phase-specific characteristics, timewise dynamics, and comprehensive variation information are readily explored. Illustrative results have demonstrated the effectiveness of the proposed method in improving process understanding and monitoring performance.

Acknowledgment

This work was supported by the Hong Kong Research Grant Council under Project 613107, the China National 973 Program (2009CB320601), and the National Natural Science Foundation of China (No. 60774068).

Literature Cited

(1) Nomikos, P.; MacGregor, J. F. Monitoring of batch processes using multi-way principal component analysis. *AIChE J.* **1994**, *40*, 1361.

(2) Nomikos, P.; MacGregor, J. F. Multivariate SPC charts for monitoring batch processes. *Technometrics* **1995**, *37*, 41.

(3) Kosanovich, K. A.; Dahl, K. S.; Piovoso, M. J. Improved process understanding using multiway principal component analysis. *Ind. Eng. Chem. Res.* **1996**, *35*, 138.

(4) Wold, S.; Kettaneh, N.; Friden, H.; Holmberg, A. Modelling and diagnosis of batch processes and analogous kinetic experiments. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 331.

(5) Louwerse, D. J.; Smilde, A. K. Multivariate statistical process control of batch processes based on three-way models. *Chem. Eng. Sci.* **2000**, *55*, 1225.

(6) Sprange, E. N. M.; Ramaker, H. J.; Westerhuis, J.; Smilde, A. Critical evaluation of approaches for on-line batch process monitoring. *Chem. Eng. Sci.* **2002**, *57*, 3979.

(7) Kourti, T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemom.* **2003**, *17*, 93.

(8) Maesschalck, R. DE.; Sanchez, F. C.; Massart, D. L.; Doherty, P.; Hailey, P. On-line monitoring of powder blending with near-infrared spectroscopy. *Appl. Spectrosc.* **1998**, *52*, 725.

(9) Workman, J. J., Jr. Review of process and non-invasive near-infrared and infrared spectroscopy. *Appl. Spectrosc. Rev.* **1999**, *34*, 1.

(10) Blanco, M.; Serrano, D. On-line monitoring and quantification of a process reaction by near-infrared spectroscopy. Catalysed esterification of butan-1-o1 by acetic acid. *Analyst* **2000**, *125*, 2059.

(11) Westerhuis, J. A.; Gurden, S. P.; Smilde, A. K. Spectroscopic monitoring of batch reactions for on-line fault detection and diagnosis. *Anal. Chem.* **2000**, *72*, 5322.

(12) Gurden, S. P.; Westerhuis, J. A.; Smilde, A. K. Monitoring of batch processes using spectroscopy. *AIChE J.* **2000**, *48*, 2283.

(13) Gabrielsson, J.; Jonsson, H.; Trygg, J.; Airiau, C.; Schmidt, B.; Escott, R. Combining process and spectroscopic data to improve batch modelling. *AIChE J.* **2005**, *52*, 3164.

(14) Zachariassen, C. B.; Larsen, J.; Berg, F. v. d.; Engelsen, S. B. Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in low methoxylated amidated pectin production. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 149.

(15) Reis, M. M.; Araújo, P. H. H.; Sayer, C.; Giudici, R. Spectroscopic on-line monitoring of reactions in dispersed medium: Chemometric challenges. *Anal. Chim. Acta* **2007**, *595*, 257.

(16) Wong, C. W. L.; Escott, R.; Martin, E.; Morris, J. The integration of spectroscopic and process data for enhanced process performance monitoring. *Can. J. Chem. Eng.* **2008**, *86*, 905.

(17) Trygg, J. Prediction and spectral profile estimation in multivariate calibration. *J. Chemom.* **2004**, *18*, 166.

(18) Benoudjit, N.; Melgani, F.; Bouzgou, H. Multiple regression systems for spectrophotometric data analysis. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 144.

(19) De Luca, M.; Oliverio, F.; Ioele, G.; Ragno, G. Multivariate calibration techniques applied to derivative spectroscopy data for the analysis of pharmaceutical mixtures. *Chemom. Intell. Lab. Syst.* **2009**, *96*, 14.

(20) Nomikos, P.; MacGregor, J. F. Multiway partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 97.

(21) Chen, J.; Wang, X. Z. A New Approach to Near-Infrared Spectral Data Analysis Using Independent Component Analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 992.

(22) Shao, X. G.; Wang, W. Z.; Hou, Y.; Cai, W. S. A new regression method based on independent component analysis. *Talanta* **2006**, *69*, 676.

(23) Westad, F. Independent component analysis and regression applied on sensory data. *J. Chemom.* **2005**, *19*, 171.

(24) Gemperline, P. J.; Yang, Y.; Bian, Z. Characterization of subcritical water oxidation with in-situ monitoring and self-modeling curve resolution. *Anal. Chim. Acta* **2003**, *485*, 73.

(25) Navea, S.; Juan, A. D.; Tauler, R. Detection and resolution of intermediate species in protein folding processes using fluorescence and circular dichroism spectroscopies and multivariate curve resolution. *Anal. Chem.* **2002**, *74*, 6031.

(26) Navea, S.; Tauler, R.; Juan, A. D. Monitoring and modeling of protein processes using mass spectrometry, circular dichroism, and multivariate curve resolution methods. *Anal. Chem.* **2006**, *78*, 4768.

(27) Hyvärinen, A. Survey on independent component analysis. *Neural Comput. Surv.* **1999**, *2*, 94.

(28) Hyvärinen, A.; Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **2000**, *13*, 411.

(29) Lee, J.-M.; Qin, S. J.; Lee, I.-B. Fault detection and diagnosis based on modified independent component analysis. *AIChE J.* **2006**, *52*, 3501.

(30) Lu, N. Y.; Gao, F. R.; Wang, F. L. A sub-PCA modeling and on-line monitoring strategy for batch processes. *AIChE J.* **2004**, *50*, 255.

(31) Zhao, C. H.; Wang, F. L.; Mao, Z. H.; Lu, N. Y.; Jia, M. X. Improved batch process monitoring and quality prediction based on multiphase statistical analysis. *Ind. Eng. Chem. Res.* **2008**, *47*, 835.

(32) Lu, N. Y.; Yao, Y.; Gao, F. R.; Wang, F. L. Two-dimensional dynamic PCA for batch process monitoring. *AIChE J.* **2005**, *51*, 3300.

(33) Yao, Y.; Gao, F. R. Batch process monitoring in score space of two-dimensional dynamic principal component analysis (PCA). *Ind. Eng. Chem. Res.* **2007**, *46*, 8033.

(34) Li, G.; Qin, S. J.; Zhou, D. H. Total PLS based contribution plots for fault diagnosis. *Acta Autom. Sin.*, manuscript accepted.

(35) Andersson, C. A. Direct orthogonalization. *Chemom. Intell. Lab. Syst.* **1999**, *47*, 51.

(36) Fearn, T. On orthogonal signal correction. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 47.

(37) Westerhuis, J. A.; Jong, S. de; Smilde, A. K. Direct orthogonal signal correction. *Chemom. Intell. Lab. Syst* **2001**, *56*, 13.

(38) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119.

(39) Trygg, J.; Wold, S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.* **2003**, *17*, 53.

(40) Yu, H.; MacGregor, J. F. Post processing methods (PLS-CCA): Simple alternatives to preprocessing methods (OSC-PLS). *Chemom. Intell. Lab. Syst.* **2004**, *73*, 199.

(41) Burnham, A. J.; Viveros, R.; MacGregor, J. F. Frameworks for latent variable multivariate regression. *J. Chemom.* **1996**, *10*, 31.

(42) Hardoon, D. R.; Szedmak, S.; Taylor, J. S. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **2004**, *16*, 2639.

(43) Kruger, U.; Qin, S. J. Canonical correlation partial least squares. In *13th IFAC Symposium on System Identification 2003, Proceedings of the Conference, 27−29 August 2003*; Elsevier Science: New York, 2003; pp 1643−1648.

(44) Ryan, T. P. *Statistical Methods for Quality Improvement*; John Wiley: New York, 1989.

(45) Norgaard, L.; Ridder, C. Rank annihilation factor analysis applied to flow injection analysis with photodiode-array detection. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 107.

(46) Bijlsma, S.; Louwerse, D. J.; Smilde, A. K. Estimating Rate Constants and Pure UV-VIS Spectra of a two-Step Reacion Using Trilinear Models. *J. Chemom.* **1999**, *13*, 311.

Received for review April 14, 2009
Revised manuscript received September 13, 2009
Accepted November 18, 2009

IE9005996