An Improved Statistical Modeling Strategy by Spectroscopy for Online Monitoring and Diagnosis of Batch Processes

Chunhui Zhao, Furong Gao, Tao Liu, and Fuli Wang

Abstract—Recently, the use of spectroscopic techniques for online process monitoring has been introduced, which are deemed to be able to provide a rich source of chemical information about operation conditions within a process system. This paper presents an improved statistical analysis and modeling strategy using spectra data for online fault detection and diagnosis of batch processes. The general principle of the proposed method is that the systematic chemical information in the successful batches can be regarded as the linear combination of some underlying unobserved and independent source spectra and the mixing coefficients as the contribution of each source to the external spectra measurements. Accordingly, it employs independent component analysis (ICA) algorithm to separate those sources and identify their time-varying effects on observed spectra throughout the batch duration and thus formulates the statistical monitoring model. Moreover, in combination with contribution plots, the actual cause of the disturbances can be diagnosed. The proposed method yields more chemical statistical meanings, results in easy model interpretation and can be readily put into online application without data estimation. Its effectiveness is successfully illustrated when applied to a case study of a two-step conversion reaction.

I. INTRODUCTION

Chemical batch processes play an important role in the production of high-quality products such as polymers, pharmaceuticals, and biochemicals. Proper monitoring and diagnosis are of great importance to ensure their operation safety and production of good quality. Batch process performance monitoring has been achieved primarily using process measurements such as temperatures, pressures, and flow rates that are measured at a frequent time interval. With the extracted information being associated with the physical parameters of the process, most statistical process monitoring systems are developed based on principal component analysis (PCA) and partial least squares (PLS) [1] –[7]. There is an increasing need for new technique development, particularly those that are easy to be realtime implemented, to improve the understanding, monitoring, and control of chemical batch processes.

In recent years, increasing attention has been paid to the statistical analysis and modelling of chemical processes through spectrometry [8]–[16]. The major advantages offered by spectroscopy are of being fast, robust, and non-destructive and are therefore an interesting and realistic alternative to classical process analytical methods. Furthermore, specific chemical properties, such as the concentration of spectroscopically active chemical species present in the reaction, changes in solvent conditions, and the presence of spectroscopically active impurities, are made available and can be directly used to identify whether the reaction is proceeding correctly. This is a major superiority compared to standard process measurements of physical conditions within a chemical process. The real-time and high-quality chemical information reveal actual chemical reaction and variation occurring in the process, which will remain in control if the process is operated successfully.

A large number of successful applications have demonstrated the effectiveness of spectroscopic techniques to a range of chemical processes, both batch and continuous. An extensive list of online spectroscopic monitoring can be found in a review by Workman et al. [17]. Most of these applications dealt with the spectra data using a calibration model for calculation of the concentration of some compounds in the reaction or for other properties that were hard to measure such as conversion, viscosity, etc., which could then be used subsequently in a monitoring scheme. Gurden et al. [10] used the spectral data much more extensively, in which, the spectroscopic measurements were directly employed to construct monitoring models and thus inspected a chemical batch reaction for fault detection and diagnosis. Generally, spectroscopic data are considered as a particular form of process data for which the process variables are a series of spectral absorbance over a range of wavelengths. Accordingly, the philosophy of monitoring chemical batches by spectroscopy is very much that of traditional statistical analysis methods with process data where the reference behaviour of systematic variation is modelled and designed based on historical data from previous normal batches, and then the behaviours of new batches are compared against it.

However, revisiting previous work, it is found that they directly adapt conventional multiway statistical scheme to work for spectroscopic analysis. Resulting from such a consideration that variable collinearity is typical in spectral absorbance over a range of wavelengths, they deal with the problem of data dimensionality and redundancy by reducing...
the number of spectral variables using feature extraction. It allows thus to shrink the original spectral data space into a subspace of smaller dimension, and then employ those underlying features for analyses. Actually, spectra measurement of a mixture is often the linear combination of the spectra of its constituent species. It would be very useful if the component spectra can be recovered from those of the mixture [18]-[20]. However, neither PCA nor PLS is designed for such a purpose. Independent component analysis (ICA) algorithm [21],[22] is able to deliver this function since it is designed to extract independent components (ICs) that constitute the observed variables. For spectroscopic analysis, it is distinctive to other methods since it is aimed at separating the spectra of the constituent components of the mixture as well as determining their effects on spectra measurement, which is called blind source signal separation process. This is clearly beneficial when all or some of components in a mixture are unknown. Chen and Wang [18] applied ICA on near-infrared spectral data, which successfully proved the effectiveness of ICA for recovering the components of interest from spectra mixture and estimating their concentrations. Moreover, as an online industrial analyzer, one main challenge of spectroscopic analysis for batch processes is the problem of data estimation. That is, the missing future spectral measurement has to be complemented during online application. Although the issue has been well solved in the analysis scheme of engineering process data, it has never been addressed in the scenario of spectroscopy.

In this work, an improved spectroscopic statistical analysis and modelling strategy is developed using ICA from a new viewpoint for online monitoring and diagnosis of batch processes. The general principle is that the spectra variation can be regarded as the linear combination of some underlying unobserved and independent sources. Accordingly, it applies ICA to the spectra measurement space to separate both source spectra and their mixing relationships, which are employed to design time-varying monitoring statistics and confidence region. It solves the disadvantages of conventional spectroscopic analysis techniques and describes how the proposed algorithm can better fit into the framework of multivariate statistical monitoring for batch processes. The application to a first-order biochemical reaction conversion process demonstrates the effectiveness of the proposed method.

II. ICA-BASED SPECTROSCOPIC ANALYSIS STRATEGY FOR ONLINE MONITORING AND DIAGNOSIS OF BATCH PROCESSES

2.1 ICA Algorithm

Independent component analysis (ICA) [21],[22] is a recently developed method in which the goal is to find the statistically independent non-Gaussian hidden factors, or as independent as possible, which constitute the observed variables through linear combinations. Such a representation seems to 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 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 both the latent components \( s \) and the demixing relationship \( W_s (R \times J) \) from only the process measurements \( x \) without any related prior knowledge, termed as the process of blind separation. Their relationship can be expressed as follows:

\[
s = W_s x \tag{1}
\]

where, \( s(R \times 1) \) denotes the independent component vector, which has unit variance: \( E(s^T s) = 1 \).

Before applying an ICA algorithm, it is assumed that process variables are pre-whitened, which is generally achieved by PCA, so that its components are uncorrelated and their variances are equivalent to unity. This whitening transformation can be expressed as

\[
z = \Lambda^{-1/2} U^T x = Q x \tag{2}
\]

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

Naturally, \( s \) can be estimated by

\[
s = Wz = WQx \tag{3}
\]

where, \( W \) is an orthogonal matrix, given that \( E(s^T s) = W E(zz^T) W^T = WW^T = I \).

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

\[
S = XW_s \tag{4}
\]

\[
A_s = (S^T S)^{-1} S^T X
\]

\[
X = S A_s + E
\]

where, \( S(J \times R) \) is the estimated independent components from the observed spectra, which actually are the estimation of the source spectra in the mixture. The mixing matrix, \( A_s (R \times N) \), actually indicate the effects of the substances on mixture spectra. Here, it readily solves the collinearity problem by guaranteeing an invertible matrix \( S^T S \) because of the mutual orthonormality of the ICs. Actually, \( S^T S \) is a diagonal matrix with identical diagonal elements. \( E(J \times N) \) is the unexplained residual.

2.2 ICA-based Monitoring and Diagnosis Strategy for Batch Processes Using Spectroscopic information

2.2.1 ICA-based Spectroscopic Modelling of Batch processes

Batch spectral data are characterized by a three-way array of measurements. In each batch run, the spectral data can be
arranged as matrix of size number of wavelengths \((J \times K)\). Then spectra information collected from similar \(I\) batches can be organized as a three-way array \(\bar{X}(I \times J \times K)\). In the present work, the batches are of equal length without special declaration.

For the modelling and monitoring of the chemical batch process, the variation between the batch runs is of main interest and this variation should be modelled. Therefore, revisiting previous work, they generally batch-wise unfold the three-way spectral data into a two-way analysis unit: \(X(I \times JK)\). Then it is analyzed and modelled in the same way as the conventional multiway PCA on engineering process data to capture the batch-wise variation and correlations between the spectra over wavelengths:

\[
\bar{X} = TP^T + E \quad \text{(5)}
\]

In this general form, \(T(I \times R)\) describes the underlying major systematic information within the normal spectral data with \(R\) PCs and \(P(JK \times R)\) reveals the correlation among spectra. \(E(I \times JK)\) contains the residual part unoccupied by the MPCA model.

There exist some disadvantages for the previous analysis strategies. On the one hand, during online application, the unknown future spectral data have to be estimated, which may distort the reliability of real-time monitoring. On the other hand, since the chemical characteristics underlying batch reaction are not consistent in time, the unified PCA loading throughout the batch operation is not enough to reveal the time-varying dynamics of underlying characteristics.

As analyzed before, spectra measurements of batch processes are more likely to the combination of chemical species present in the reaction. Further, along time direction, different source species impose varying effects on the mixture spectra. For example, the first spectrum mainly equals the spectrum of the pure reactant; after a few minutes, the spectrum of the intermediate becomes visible and then slowly disappears again; finally, the spectrum is determined by that of the end-product. The alternation of domimative source species actually is externally reflected by the changes of mixing coefficients. The larger the magnitude of mixing coefficients, the more important the estimated ICs. Moreover, the values of a mixing matrix at a specific time for all batches satisfy a Gaussian distribution since they describe the normal variation over different batches. Based on the above cognition and analysis, the mixing coefficients are employed in the proposed algorithm to develop the reference model and check whether the reaction is proceeding smoothly.

The procedure for this new method is presented in Figure 1.

First, the three-way batch spectra data is organized as \(\bar{X}(I \times JK)\), in which, each column reveals the spectrum from each batch at each time.

Then ICA is run on them, which can extract all the underlying sources throughout the process duration from an overall viewpoint and meanwhile takes into account the interaction between different samples along both time and batch directions. The ICA model is designed as follows:

\[
S(J \times R) = XW
\]

\[
A(R \times IK) = \left( S^T S \right)^{-1} S^T X \quad \text{(6)}
\]

\[
E(J \times IK) = X - SA = \left( I - S \left( S^T S \right)^{-1} S^T \right) X
\]

where, the sources, \(S(J \times R)\), summarize all spectra once emerged in the chemical reaction; and mixing coefficients, \(A(R \times IK)\), indicate their corresponding roles played in mixture spectra along time direction.

Accordingly, the time-slice mixing matrix, \(A(I \times R)\), and the time-varying residual, \(E(I \times J)\), are separated, from which, we can be informed the time-varying effects of underlying species on spectra measurements and the unmodeled chemical information at each time. They will be employed for the subsequent calculation of monitoring statistics.

**2.2.2 Monitoring Statistics**

To reveal the time-varying chemical information for online application, monitoring statistics are developed at each time. The separated time-slice mixing matrix, \(A(I \times R)\), and residual, \(E(I \times J)\), cover and explain the normal batch-wise variation at each time. Thus, they are employed to construct...
two online monitoring statistics: $D^2$ in systematic subspace and the squared prediction error (SPE) in residual subspace.

$D^2$ is defined as follows and its confidence limits can be obtained from the indicated $F$ -distribution with a significance factor:

$$D_{ik}^2 = (a_{ik} - \bar{a}_k)^T M_k^{-1} (a_{ik} - \bar{a}_k) - \frac{R(I^2 - 1)}{I(I - R)} F_{r,I-R,a}$$

(7)

where, $a_{ik} (R \times 1)$ are the mixing vector corresponding to $i$th batch at time $k$ ; $\bar{a}_k (R \times 1)$ denotes the mean vector of time-slice mixing matrix $A_k (I \times R)$ . $M_k (R \times R)$ is the covariance matrix of $A_k (I \times R)$.

The SPE is a measure of the lack of fit with the established model, and is defined as the sum of the squares of the errors at time $k$:

$$\text{SPE}_k = e_{ik}^T e_{ik}$$

(8)

where, $e_{ik}$ is the residual vector of the $i$th batch at time $k$ which actually can be calculated by

$$e_{ik} = x_{ik} - Sa_{ik} = (I - S(S^T S)^{-1} S^T) x_{ik}.$$  

The control limit of SPE can be approximated by a weighted $\chi^2$ distribution:

$$\text{SPE}_k \sim g_k^2 \chi_{h_k,a}^2$$

(9)

where, $g_k = v_k / 2m_k$ and $h_k = 2(m_k)^2 / v_k$, in which, $m_k$ is the average of all the SPE values at time $k$ calculated in eq (8), and $v_k$ is the corresponding variance.

The above two monitoring statistics provide reference standard for checking the status of a new batch. When a new measurement sample, $x_{\text{new}} (J \times 1)$, is available, indicated by the current process time, it is analyzed using the designed ICA model and the monitoring statistics are calculated as follows:

$$a_{\text{new}} = (S^T S)^{-1} S^T x_{\text{new}}$$

$$e_{\text{new}} = x_{\text{new}} - Sa_{\text{new}} = (I - S(S^T S)^{-1} S^T) x_{\text{new}}$$

(10)

$$D_{\text{new}}^2 = (a_{\text{new}} - \bar{a}_k)^T M_k^{-1} (a_{\text{new}} - \bar{a}_k)$$

$$\text{SPE}_{\text{new}} = e_{\text{new}}^T e_{\text{new}}$$

Process monitoring is conducted by continuously comparing both statistics with the predetermined control limits. If they stay well within the predefined normal region, the chemical batch process can be deemed to be subject to normal reaction. Otherwise the statistics will go beyond the control limits responding to abnormal variations.

2.2.3 Contribution Plot for Fault Diagnosis

Once the unusual events are detected unoccupied in the reference batches, it is often followed by fault diagnosis, which tries to find out what is wrong with the process, or what causes the failure of chemical reaction. Here, the idea of contribution plot [23],[24] is inherited to isolate the responsible wavelengths to the violated monitoring indices at each time.

On the basis of eq 7, the contribution to the $D^2$ -statistic shared by the spectrum at each wavelength can be defined by investigating the individual effect of each single-wavelength spectrum on it:

$$C_{D_{ij},j} = (a_{\text{new}} - \bar{a}_k)^T M_k^{-1} \left( v_j x_{\text{new},j} - \bar{a}_{ki,j} \right)$$

(11)

where, vector $v_j$ is the $j$th column of the matrix $(S^T S)^{-1} S^T$, $x_{\text{new},j}$ is the $j$th element of $x_{\text{new}} (J \times 1)$ and $\bar{a}_{ki,j}$ is the $j$th element of mean vector $\bar{a}_k (R \times 1)$, which all correspond to the concerned spectrum at the $j$th wavelength.

Moreover, the contribution to SPE is also captured based on the same fact that the contributions summed over all $J$ wavelengths equal the SPE value of the new batch:

$$C_{\text{SPE}_{ij},j} = e_{\text{new},j}^2$$

(12)

where, $e_{\text{new},j}$ is the residual of spectrum at the $j$th wavelength.

III. APPLICATION TO A TWO-STEP CHEMICAL BATCH PROCESS

In this section, the proposed monitoring method is applied to on-line fault detection and diagnosis of a biochemical two-step conversion reaction of 3-chlorophenylhydrazonopropane dinitrile (A) with 2-mercaptoethanol (B). During the reaction process, an intermediate adduct (C) is formed and then hydrolyzed to give the main product 3-chlorophenylhydrazonocyanamidamide (D) and ethylene sulphide (E):

$$A+B \rightarrow C \rightarrow D+E$$

(13)

The reaction takes place in a quartz cuvette using a reactant volume of 2.5 mL with a water bath and thermocouple to maintain a temperature of 25 °C. From the time of reaction initiation, a spectrum with a wavelength range of 300-500 nm at a resolution of 1nm was recorded every 20 s for a total run time of 45 min. A set of 37 batches, both successful and unsuccessful, are collected, in which, 27 successful batches are used to develop the monitoring system; and 3 normal and 7 abnormal batches are monitored online to validate how well the monitoring system can indicate the status of chemical reaction.

Using the modelling strategy described in Section 2, the underlying independent sources are extracted from the spectra measurement and then based on mixing coefficients monitoring statistical indices are designed as well as their normal regions. In the present work, 5 ICs are retained and the corresponding mixing relationships are shown in Figure 2 taking example for the first four ICs. The changing trend clearly reveals an obvious two-step chemical reaction since they yield different effects on spectra measurement.
jointly indicate that the batch behaves free of any process abnormality throughout the duration. For the fault case, pH disturbance is introduced by adding 10 μL NaOH with concentration 0.1020 mol/L to the reaction mixture. From the fault detection result shown in Figure 4, the upset can be quickly picked up by synthetically analyzing the monitoring charts, where the statistic values obviously go beyond the control limits almost immediately once the faults occur. Therefore, the monitoring models can effectively and realtime capture the abnormal variation occurred during the chemical reaction process with no data estimation cost. Besides its comprehensive analysis of underlying source spectra, the better performance may also benefit from the introduction of time-varying covariances.

Figure 3 shows the online monitoring results for one successful batch. From the plot it can be seen that both monitoring indices stay well within the control bounds, which

due to the interference of a process abnormality. For the fault case, pH disturbance is introduced by adding 10 μL NaOH with concentration 0.1020 mol/L to the reaction mixture. From the fault detection result shown in Figure 4, the upset can be quickly picked up by synthetically analyzing the monitoring charts, where the statistic values obviously go beyond the control limits almost immediately once the faults occur. Therefore, the monitoring models can effectively and realtime capture the abnormal variation occurred during the chemical reaction process with no data estimation cost. Besides its comprehensive analysis of underlying source spectra, the better performance may also benefit from the introduction of time-varying covariances.

Figure 3 shows the online monitoring results for one successful batch. From the plot it can be seen that both monitoring indices stay well within the control bounds, which

IV. CONCLUSION

In this paper, an improved spectroscopic statistical analysis strategy is presented for fault detection and diagnosis of a chemical batch reaction. By extracting those sources underlying spectra measurement and identifying their combination relationship using ICA algorithm, monitoring statistics and confidence regions are developed. It overcomes some advantages of the conventional spectroscopy monitoring methods and can be readily put into online application without data estimation cost. Simulation examples have shown the effectiveness of the proposed method with fast and accurate fault detection performance. This proposed algorithm provides an analysis tool of batch processes for more informative statistical explanations and better chemically interpretable characteristics by making
better use of spectroscopic information. Further research is thus suggested.

ACKNOWLEDGMENT

Many thanks are due to Johan A. Westerhuis in Department of Chemical Engineering, University of Amsterdam for providing the related spectral data. The work was supported in part by Hong Kong Research Grant Council under project number 613107 and National Natural Science Foundation of China (No. 60774068).

REFERENCES


