Modeling and Performance Monitoring of Multivariate Multimodal Processes

Thiago Feital
Programa de Engenharia Química, COPPE, Universidade Federal do Rio de Janeiro, Brazil

Uwe Kruger
Dept. of Mechanical & Industrial Engineering, Sultan Qaboos University, Muscat, Oman

Julio Dutra, José Carlos Pinto, and Enrique Luis Lima
Programa de Engenharia Química, COPPE, Universidade Federal do Rio de Janeiro, Brazil

DOI 10.1002/aic.13953
Published online November 26, 2012 in Wiley Online Library (wileyonlinelibrary.com).

A multimodal modeling and monitoring approach based on maximum likelihood principal component analysis and a component-wise identification of operating modes are presented. Analyzing each principal component individually allows separating components describing the variation within the individual modes from those capturing variation which the modes commonly share. On the basis of the former set, a Gaussian mixture model produces a statistical fingerprint that describes the production modes. The advantage of the component-wise analysis is a simple identification of the mixture model parameters, which does not rely on the computationally cumbersome expectation maximization. The proposed method diagnoses abnormal process conditions by defining statistics relating to the components describing (1) between-cluster variation, (2) within cluster variation, and (3) model residuals. The article demonstrates the benefits of this approach over existing work by an application to a continuous stirred tank reactor (CSTR) simulator and the analysis of recorded data from a furnace and a chemical reaction process. © 2012 American Institute of Chemical Engineers *AIChE J,* 59: 1557–1569, 2013
Keywords: multimodal processes, Gaussian mixture models, component-wise identification, within- and between-cluster variation, maximum likelihood principal component analysis

Introduction

For large-scale production systems in the chemical industry, the presence of abnormal situations may result in the release of toxic material to the environment, fires and explosions, and can have significant economic implications caused by fines, loss of production, and damage to the operation units.[1] Data-driven process monitoring presents a potential solution to identify the presence of abnormal operating conditions. Among existing monitoring approaches, discussed in the research literature over the past few decades, the component technology collectively referred to as multivariate statistical process control (MSPC) has gained considerable attention resulting from its conceptual simplicity, practical usefulness, and the interpretability of monitoring charts by plant personnel.[2]

MSPC uses historical process data sets describing normal operation condition (NOC) to generate rules and/or tests for fault detection and isolation (FDI). MSPC-based FDI approaches have been discussed in the research literature from the early 1990s and, for example, include Refs. 2–8. Despite numerous successful application studies involving MSPC technology, the data models used assume a single steady-state operating condition. Industrial process systems in the chemical industry, however, are often characterized by multiple operating modes, caused by feedstock, product specifications, set-points, and/or manufacturing strategies.[9,10] Consequently, statistical approaches that assume a single operation conditions are only of limited value.

Among the multimodal approaches recently proposed in literature and discussed in next section, there are a number of practically and conceptually important issues that have not been adequately addressed. These mainly relate to the modeling task and involve the determination and description of multimodal variables and how to identify the number of modes. Additionally, the incorporation of a multimodal model into an on-line FDI scheme is also an issue that is also not adequately discussed in the literature.

To address these issues, the article combines previous work on multivariate multimodal process monitoring by integrating maximum likelihood principal component analysis (MLPCA) and a component-wise analysis approach that is developed here. First, the application of MLPCA estimates the source variables describing the different operating conditions with respect to the following data structure

$$\mathbf{z}(k) = \mathbf{A}\mathbf{s}(k) + \mathbf{e}(k), \tag{1}$$

where k is a sampling index, $\mathbf{z} \in \mathbb{R}^N$ is a vector storing recorded process variables, $\mathbf{s} \in \mathbb{R}^n$ is a vector of n source signals representing common cause variation $n < N$, $\mathbf{e} \in \mathbb{R}^N$ is an error vector describing, for example, measurement

Correspondence concerning this article should be addressed to U. Kruger at uwekruger@squ.edu.om.

uncertainty, and $\mathbf{A} \in \mathbb{R}^{N \times n}$ is a parameter matrix. The source and error variables are assumed here to follow a Gaussian distribution for each mode, that is, $\mathbf{s} \sim \mathcal{N}\left\{\bar{\mathbf{s}}^{(i)}, \mathbf{S}_{ss}^{(i)}\right\}$ with $1 \leq i \leq M$ being the ith mode, $\mathbf{e} \sim \mathcal{N}\{\mathbf{0}, \mathbf{S}_{ee}\}$ and $\mathbf{z} \sim \mathcal{N}\left\{\mathbf{A}\bar{\mathbf{s}}^{(i)}, \mathbf{A}\mathbf{S}_{ss}^{(i)}\mathbf{A}^T + \mathbf{S}_{ee}\right\}$. Throughout this article, the notation $\bar{\cdot}$, for example, $\bar{\mathbf{z}}$, refers to a mean vector and $\mathbf{S}_{\cdot\cdot}$, for example, $\mathbf{S}_{zz}$, denotes a covariance matrix. For simplicity, Eq. 1 assumes that the mean vector for the data vector $\mathbf{z}$ is determined by the source signals. This is not a restriction of generality as it is straightforward to incorporate an additional mean vector, such that $\mathbf{z}(k) = \mathbf{A}\mathbf{s}(k) + \mathbf{e}(k) + \bar{\mathbf{z}}$. The article also elaborates special cases where $n = N = m$ and $n = N > m$.

This article proposes a recursive scheme for identifying modes in which each principal component is analyzed individually. The scheme tests whether a sample (1) belongs to one of the existing operating modes, (2) belongs to a new operating mode, or (3) is abnormal, for example, an outlier. The component-wise identification allows separating the principal components, and hence the source signals, into those describing the different modes and the ones that capture variation that the individual production modes commonly share. Such a detailed modeling scheme has not yet been proposed in the literature for multimodal process systems to the best of the authors' knowledge.

A significant benefit of (1) the component-wise analysis and (2) the separation of the principal components is the identification of a Gaussian mixture model (GMM) for the components describing the individual modes. Besides reducing the dimensionality, the GMM parameters can be obtained with no significant effort, as opposed to the application of expectation maximization (EM), which is computationally expensive and may yield numerical problems in determining the unknown parameters.

The article is organized as follows. The next section provides a discussion and an analysis of recent work on monitoring multimodal processes. The two sections that follow present a brief summary of the relevant component technology and introduce the proposed multivariate multimodal approach. To outline the benefits of the proposed approach, the article then contrasts the working of the proposed technique with a conventional MLPCA-based monitoring approach and the multimodal method discussed in Ref. 9. This comparison is presented in two separate sections and includes the application to a CSTR simulator and recorded data from a chemical reaction process. The article also presents an example involving recorded data from an industrial furnace in a separate section. The final section then presents a concluding summary of this article.

Previous Work on Modeling and Monitoring Multimodal Processes

This section reviews and analyzes existing work on the modeling and monitoring of multimodal systems. On the basis of this analysis, the motivation for the proposed multimodal approach is given at the end of the section.

Mixture of component or Gaussian models

Reference 11 proposed a mixture of principal component analysis (PCA) models for fault detection. This approach uses a heuristic smoothing clustering algorithm to automatically determine the number of clusters and an updating procedure to encompass new events into the mixture model. Also in an *ad hoc* formulation, Ref. 12 proposed multiple partial least-squares models for multimodal process monitoring. The number of clusters is assumed to be known *a priori* in this case. More recently, Ref. 13 developed an Adjoined PCA approach, where a mixture of PCA models are adjoined, allowing a smooth transition between two models. Another approach dealing with model transitions can be found in Ref. 14.

Regarding probabilistic models, Ref. 15 formulated a probabilistic PCA mixture model, whose parameters are determined simultaneously. Along the same lines, Ref. 16 formulated the MLPCA mixture model to use as a process monitoring tool. Their approach also relies on local models, which are generated simultaneously with the help of a learning algorithm. More recently, Ge and Song[17] proposed a maximum likelihood factor analysis mixture model and used it in a process monitoring context.

Apart from the multivariate approaches, Yu and Qin[9] developed a multimodal monitoring approach using a GMM based on the recorded variables and a Bayesian index. The number of modes is automatically determined along with the other parameters and the operating point transition is performed intrinsically by the Bayesian index using the probabilities of the clusters.

Common structures

Reference 10 introduced an extension of PCA, assuming a common principal subspace of all clusters. This approach, however, requires *a priori* knowledge including the number of modes and the respective samples of each mode. A similar approach was proposed by Maestri et al.[18] based on a robust algorithm to determine the common covariance structure for all modes.

Hybrid approaches

Reference 19 proposed a multivariate multimodal process monitoring approach using PCA and GMM. Each principal component is assumed to describe the different modes. Moreover, this approach considers a common error structure for every cluster. More recently, Chen and Zhang[20] introduced a similar monitoring approach for multimodal batch processes. This technique applies multiway PCA on the process data and uses the principal components along with the logarithm of the squared prediction error (log-SPE) as the set of variables to generate a GMM.

Alternative approaches

These approaches include Yu,[21] which uses PCA or independent component analysis (ICA) in conjunction with Hidden Markov models, and Refs. 22 and 23 which propose the use of local Least-Squares Support Vector Regression and ICA–PCA. Ge and Song[24] proposed the combined utilization of Fuzzy C-mean and ICA–PCA. Each of these alternative approaches may also be categorizes as hybrid ones and use advanced techniques to identify a process model.

Analyses of existing approaches

The first type of multimodal approaches, mixture of component or Gaussian models, applied directly to all process variables, may require an excessive number of parameters which may lead to redundant information shared by the individual clusters. Examples of this include the mass balances and other constraints that are present in each process operating point. Using a common error structure, however, is a

practically reasonable assumption. Without such an assumption, the application of a mixture of models increases the complexity of the monitoring approach and reduces the interpretability. For example, Section 2 in Camacho et al.[25] presents a motivating example which illustrates that a simple blending process, depending on the nominal flow of each stream, may form different clusters along the direction of the principal components.

Different from mixtures of Gaussian models, the assumption for the second type, common structures, is that the physical rules governing the process are identical and, consequently, the clusters share the same covariance structure. This assumption, however, can only be true if the variable interrelationships are identical for each operating point. For example, chemical reactions and transport phenomena may be different for different operating points. Furthermore, different operating points may have different connections (recycles and purges) and control structures. This, in turn, introduces diversity among the operating modes and produces different covariance structures for the clusters.

The motivation for hybrid approaches centers on the deficiencies of the former two methods. As they may be either too complex or too simplistic, a reasonable alternative is to combine both approaches, which yield (1) a multivariate methods to identify a common error structure and (2) multimodal models to characterize different operating modes. The hybrid approaches rely, mainly, on PCA to develop component models that extract estimates of the source signals describing the production modes. The operating points are assumed to be known *a priori* or identifiable by a clustering method and the extracted source signals are all assumed to have multimodal behavior, which this article shows may not be the case. According to Eq. 1, the source and error signals are assumed to follow Gaussian distribution functions for a specific operation mode. Relying on PCA, however, may not yield a consistent estimation of the model and residual subspaces.[26]

Finally, integrating advanced modeling methods, such as support vector regression, Fuzzy C mean, or ICA, into hybrid approaches increases the complexity of the modeling task. For practical reasons, however, it is important to note that the monitoring approaches for multimodal systems must be implemented on-line and understood by process operators. In this regard, the practical success of the on-line monitoring scheme is directly related to the ability of the plant operators in applying the component technology. In addition, the application studies in this article demonstrate that the assumption of Gaussian process variables can provide sufficiently accurate monitoring models.

Motivation for proposed approach

The motivation for the work reported in this article relies on the following issues, which have not been addressed in the research literature. More precisely, this article proposes an approach that simultaneously:

- deals with a general statistical description of the error variables, as PCA relies on the assumption that $\mathbf{S}_{ee} = \sigma_e^2\mathbf{I}$ with σ_e^2 being the noise variance for all recorded process variables;
- takes advantage of a reduced dimensional data representation and can be directly applied for unknown operating modes;
- determines the number of multimodal variables, or source signals;
- relies on conventional MSPC techniques for modeling and monitoring multimodal processes; and
- separately monitors component describing different operating modes and variation that each of the modes commonly share.

Preliminaries

This section presents a brief summary of PCA and GMM and introduces the required nomenclature for the remainder of this article.

Principal component analysis

PCA determines a model and a complementary residuals subspace on the basis of the estimated data covariance matrix $\mathbf{S}_{zz} = \frac{1}{K-1}\sum_{k=1}^{K}(\mathbf{z}(k) - \hat{\bar{\mathbf{z}}})\,(\mathbf{z}(k) - \hat{\bar{\mathbf{z}}})^T$, with $\hat{\bar{\mathbf{z}}}$ being the estimated mean vector of $\mathbf{z}$.[27] This method decomposes a given and scaled data matrix $\mathbf{Z} \in \mathbb{R}^{K\times N}$ into a score and a loading matrix $\mathbf{T} \in \mathbb{R}^{K\times n}$ and $\mathbf{P} \in \mathbb{R}^{N\times n}$, respectively, and a residual matrix $\mathbf{F} \in \mathbb{R}^{K\times N}$, where K is the number of reference samples and n is referred to as the number of principal components. Equation 2 shows the decomposition for the kth sample and the determination of the score vector in more detail

$$\mathbf{z}(k) = \mathbf{P}\mathbf{t}(k) + \mathbf{f}(k) \quad \mathbf{t}(k) = \mathbf{P}^T\mathbf{z}(k), \tag{2}$$

where $\mathbf{t} \in \mathbb{R}^n$ and $\mathbf{f} \in \mathbb{R}^N$. Using MLPCA, Feital et al.[26] showed how to estimate $\mathbf{P}$ under the assumption that the error covariance matrix in Eq. 1 $\mathbf{S}_{ee}$, is nonisotropic. Defining $\mathbf{\Lambda}$ as a diagonal matrix storing the variances of the score variables, Eq. 3 defines the Hotelling's T^2 and Q statistics, which are typically used for process monitoring applications along with control limits that can be obtained from an F and approximated by a χ^2 distribution, respectively[2]

$$T^2(k) = \mathbf{t}^T(k)\mathbf{\Lambda}\mathbf{t}(k) \quad Q(k) = \mathbf{f}^T(k)\mathbf{f}(k), \tag{3}$$

Gaussian mixture model

This is a probabilistic model represented by a weighted average of Gaussian probability density functions (PDFs) describing a set of production modes

$$f(\mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\mu}_z, \mathbf{\Sigma}_z) = \sum_{i=1}^{M}\boldsymbol{\pi}_i\mathcal{N}\left\{\mathbf{z}|\bar{\mathbf{z}}^{(i)}, \mathbf{S}_{zz}^{(i)}\right\}. \tag{4}$$

Here, $\boldsymbol{\pi} \in \mathbb{R}^M$, $\boldsymbol{\mu}_Z \in \mathbb{R}^{N\times M}$, and $\mathbf{\Sigma}_Z \in \mathbb{R}^{N\times N\times M}$ are the cluster probabilities (weighting parameters for each of the M modes), a matrix storing the variable means as column vectors and an augmented arrangement storing the covariance matrices for each cluster, respectively. The parameters of the PDF $f(\cdot)$ can be obtained by an EM algorithm.[28] The algorithm proposed in this article relies on a two-step clustering approach that is used to estimate the GMM. It is important to note that the proposed approach does not suffer from the computational and particularly numerical problems of EM algorithms.[28]

Statistical-Based Modeling and Monitoring of Multimodal Systems

This section introduces the proposed modeling and monitoring technique for multivariate multimodal process

systems. The underlying assumption is that each individual mode not only possesses a specific covariance structure but these production modes also share common covariance information. Geometrically, the NOC region for the ith operating mode is a hyperellipsoid and samples whose projections fall inside this NOC region describe in-statistical-control behavior. In contrast, if the projection of samples fall outside the NOC region, they are assumed to show out-of-statistical-control or abnormal operation condition (AOC) behavior with respect to the ith operating mode.

Process modeling

To construct a NOC region from recorded reference data of the process, the proposed technique relies on Eq. 1. Dividing the extracted source signals $\mathbf{s}$, into two variable sets that describe the between-cluster $\mathbf{s}_1 \in \mathbb{R}^m$, $m \leq n$ and the within-cluster variation $\mathbf{s}_2 \in \mathbb{R}^{n-m}$ yields the following data structure

$$\mathbf{s} = \begin{pmatrix} s_1 & \cdots & s_m & s_{m+1} & \cdots & s_n \end{pmatrix}^T = \begin{pmatrix} \mathbf{s}_1 & \mathbf{s}_2 \end{pmatrix}^T; \quad (5a)$$

$$\mathbf{s}_1 \sim \sum_{i=1}^{M} \pi_i \mathcal{N}\left\{\bar{\mathbf{s}}_1^{(i)}, \mathbf{S}_{ss}^{(i)}\right\} \text{ and } \mathbf{s}_2 \sim \mathcal{N}\{0, \mathbf{S}_{s_2 s_2}\} \quad (5b)$$

$$\mathbf{z}(k) = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \begin{pmatrix} \mathbf{s}_1(k) \\ \mathbf{s}_2(k) \end{pmatrix} + \mathbf{e}(k) \quad (5c)$$

More precisely, the PCA model subspace is divided into an m dimensional subspace covering the covariance structure of the between-cluster variation and a complementary $n - m$ dimensional subspace describing a common within-cluster covariance structure.

Applying MLPCA allows extracting the source signals in the noise free case, $\mathbf{e} = \mathbf{0}$ in Eq. 1, and estimating them for $\mathbf{e} \sim \mathcal{N}\{\mathbf{0}, \mathbf{S}_{ee}\}$ up to a similarity transformation.[2] With respect to the data structure in Eq. 1, Feital et al.[26] introduced a statically based test for determining n. Following the assumptions of the source signals in Eqs. 5a–5b, m can simply be determined by inspecting which of the score variables identify more than one mode. More precisely, components for which $M = 1$ follow, according to Eq. 5b, a Gaussian PDF.

After determining m, a GMM can be determined, which gives rise to the estimation of π, $\mu_{s_1} \in \mathbb{R}^{m \times M}$ and $\sum_{s_1} \in \mathbb{R}^{m \times m \times M}$ with respect to Eq. 4. Different to existing work, this article proposes a clustering procedure that includes (1) a univariate parameter estimation and (2) a multivariate parameters estimation. As briefly touched on in the preceding discussion, this approach is not subjected to the problems of using the EM algorithm, which is used to determine GMMs in Refs. 16, 17, 19, and 20. The next subsection introduces the proposed approach and discusses how to use it for online monitoring.

Univariate Estimation As the process is assumed to operate at the ith operating point for some time, each of the score variables follows a Gaussian distribution according to Eq. 5

$$t^{(i,j)}(k) = \mathbf{p}_j^T \mathbf{z}(k) \sim \mathcal{N}\{\bar{t}^{i,j}, \sigma^{(i,j)}\}. \quad (6)$$

Here, $\bar{t}^{(i,j)}$ and $\sigma^{(i,j)}$ describe the mean and variance of the jth score variable $1 \leq j \leq n$, respectively. Following from Eqs. 4 and 5a, it is important to note that the source signal can be described by a GMM in a global (for all M clusters) context and follow a Gaussian PDF in local context (for the ith operation point). The second assumption is that the period at which the process operates at the ith operating point is sufficiently long for estimating $\bar{t}^{(i,j)}$ and $\sigma^{(i,j)}$.

The third assumption is that the data describing the ith operating point is fault-free. This assumption guarantees that all the estimated parameters relate to a NOC only. These three assumptions give rise to the following hypothesis for testing whether the jth score variable describes the ith operating condition

$$H_0 \ : \ \bar{t}^{(i,j)} - c_\alpha \sqrt{\sigma^{(i,j)}} \leq \mathbf{p}_j^T \mathbf{z}(k) \leq \bar{t}^{(i,j)} + c_\alpha \sqrt{\sigma^{(i,j)}}; \quad (7a)$$

and

$$H_1 \ : \ |\mathbf{p}_j^T \mathbf{z}(k)| > \bar{t}^{i,j} + c_\alpha \sqrt{\sigma^{(i,j)}}. \quad (7b)$$

Here, H_0 is the null hypothesis (the kth sample belongs to the ith operating point), H_1 is the alternative hypothesis (this sample does not belong to the ith cluster) and c_α is control limit for a standard Gaussian distribution of significance α. If the null hypothesis is accepted, the kth sample belongs to the ith mode. Rejecting the null hypothesis implies that the process operates at a different operating or the sample is an outlier.

If H_0 is rejected, each of the known operating condition needs to be tested on the basis of Eqs. 7a and 7b. If none of the identified clusters match the kth sample, an additional mode is included and $M = M + 1$. If the kth sample matches one of the M operating conditions, however, the sample is added to the set of samples describing this mode and the mean and variance of the score variables for this mode are recursively updated. This procedure, detailed in Figure 1, gives rise to a recursive determination of the modes using one principal component at the time. Components which only have $M > 1$ modes describe between-cluster variation and are stored in $\mathbf{t}_1$. Conversely, components for which $M = 1$ capture within-cluster variation and are, consequently, stored in $\mathbf{t}_2$. Principal components that correspond to between-cluster variation are expected to be associated with larger eigenvalues; whereas those corresponding to within-cluster variation have, typically, smaller eigenvalues.

It is important to note that this approach does not directly address transient changes between operating points. As the control limits are adapted as additional samples are added to the ith cluster, slow transitions may lead to significant overlaps among the steady-state operating conditions. The instances of such changes in operating conditions are known by inspecting the production records and transient changes can be modeled using the approach proposed in Ref. 29. Based on this reference, the incorporation of such transient changes, however, is straightforward if a number of such transients are available in the data records. To illustrate the working of the proposed multimodal technique, the issue of significant transient changes is not considered here. It should be noted, however, that the proposed approach can deal with the following two types of transients: (1) fast transients, which only cover a few samples between two operating modes that can be removed from the computation of the adaptive control limits for the modes concerned, and (2) gradual and optimized transients, for example, by supervisory controllers, which may give rise to several small clusters between the clusters describing the two individual modes.

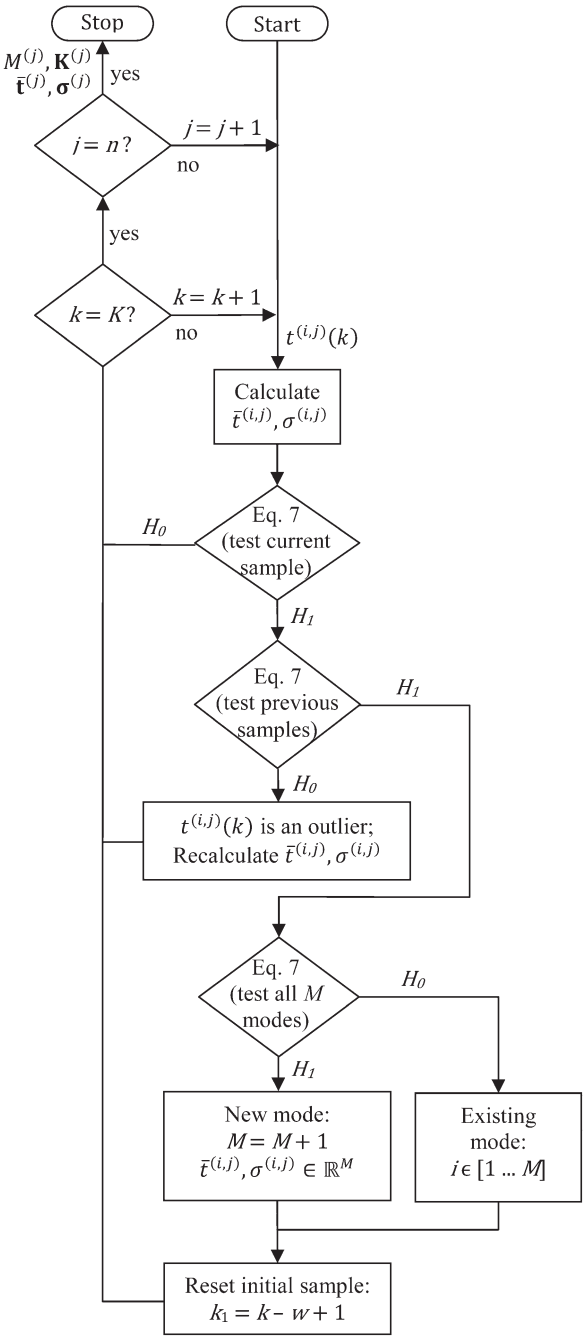

Figure 1. Flowchart detailing the working of the introduced component-wise clustering algorithm.

Multivariate Estimation After determining the M distinct operating conditions using the univariate step, the next modeling step is to determine the GMM parameters for the m score variables in $\mathbf{t}_1$. This is one of the benefits of the proposed approach, as it relies on $m \leq n < N$ variables, which simplifies the modeling task at hand. Following the univariate estimation step, the mean vector for each of the modes, $\bar{\mathbf{t}}_1^{(i)}$, is available and the values for each of the m score variables are divided with respect to the individual clusters, identifying the $K^{(i)}$ samples of the ith cluster. This allows estimating the covariance matrices for each mode, $\mathbf{S}_{t_1 t_1}{}^{(i)}$. The remaining parameters of the GMM model, π_i, can now be determined as the ratio of the number of samples of the ith cluster, $K^{(i)}$, over the total number of samples K. It is straightforward to show that the resultant GMM model presents an estimation of the PDF for the m score variables describing the M operating modes.

On-line statistical process monitoring

After determining each of the M operating condition and a GMM model describing the PDF of the m score variables, on-line process monitoring entails the following three steps:

1. Identify the current operating point;
2. Determine whether the process operates in-statistical-control or out-of-statistical-control; and
3. Analyze abnormal process behavior if the process is out-of-statistical-control.

These steps are described in the next three subsections.

Step 1: Identify the Current Operating Point Most processes operate under closed-loop control and the set points for supervisory controllers characterize the current operating mode. Abbreviating the operating condition by *op*, common cause variation, described by the source variables stored in $\mathbf{s}$, is introduced to the process by measured and unmeasured disturbances. Mathematically, the operating mode is a function of the set point for the supervisory controllers, defined by $\mathbf{y}^{sp}$, and a variable $\mathbf{d}$ set describing the disturbances

$$op = f(\mathbf{y}^{sp}, \mathbf{d}) \tag{8}$$

Note that the role of the regulatory controllers is to counter the impact of disturbances, such that the process operates at the current operating condition or move to another operation point if required. As outlined in the preceding discussion, however, transient changes are not considered in this article but the extension of the proposed method to monitor such transitions follows from the discussion in Kourti[29] and is straightforward.

To determine the current operating point in cases of open-loop processes or unavailable set points, a maximum likelihood criteria can be considered, based on a moving window approach for a window length w.[30] For the kth sample, the set of sample indices for the corresponding data window include $\{k - w + 1, k - w + 2, \ldots, k - 1, k\}$. Using this data window, the current operating point can be identified through the following maximum likelihood objective function for $1 \leq i \leq M$:

$$i = \arg\max_i \prod_{j=k-w+1}^{k} f(\mathbf{t}(j)|\pi_i, \bar{\mathbf{t}}_1^{(i)}, \mathbf{S}_{t_1,t_1}^{(i)}). \tag{9}$$

Here, $f(\cdot)$ is shown in Eqs. 4 and 5b, π_i, $\bar{\mathbf{t}}_1^{(i)}$ and $\mathbf{S}_{t_1 t1}{}^{(i)}$ are the estimates for the GMM model, the mean vector and the covariance matrix, respectively. The integer i, for which Eq. 9 assumes a maximum, is then the current operating condition, that is, $i = op$.

Step 2: Process Monitoring To monitor the operating condition, *op*, and to detect the presence of an AOC, this second step is based on the construction of three monitoring statistics. The first two statistics rely on the score variables $\mathbf{t}_1$ and $\mathbf{t}_2$, defined here as the D and T^2 statistics, while the third one is a residual Q statistic constructed from the MLPCA model residuals

$$D(k) = (\mathbf{t}_1^{(op)}(k) - \bar{\mathbf{t}}_1^{(op)})^T \, \mathbf{S}_{t_1 t_2}^{(op)^{-1}} \, (\mathbf{t}_1^{(op)}(k) - \bar{\mathbf{t}}_1^{(op)}) \tag{10a}$$

$$T^2(k) = \mathbf{t}_2^T(k) \boldsymbol{\Lambda}_2 \mathbf{t}_2(k). \tag{10b}$$

Here, $\boldsymbol{\Lambda}_2$ is a diagonal matrix storing the eigenvalues associated with the $n - m$ score variables describing within

cluster variation. The residual Q statistic is defined in Eq. 3. It is important to note that if $m = n$, all of the score variables describe between-cluster variation and no common within-cluster variation is present. In this case, only the D statistic can be established.

Both statistics in Eq. 10 follow an F distribution, with m and $K - m$ degrees for freedom for the D statistic and $n - m$ and $K - n + m$ degrees of freedom for the T^2 statistic. The distribution function for the Q statistic can be approximated by a scaled χ^2 distribution.[31,32] Testing the null hypothesis that the process is in-statistical-control with respect to *op* if a new sample becomes available is based on the following hypothesis tests

$$D(k) \le D_\alpha \qquad T^2(k) \le T^2_\alpha \qquad Q(k) \le Q_\alpha, \tag{11}$$

where D_α, T^2_α, and Q_α are the control limits of the D, the T^2, and the Q statistic for a significance of α, respectively. If the null hypothesis for the D statistic is rejected, it needs to be tested whether the new sample matches a different operating condition, that is, whether the process has shifted to a different operating condition such that $i \neq op$. If it does not match any of the M operating condition, the process is out-of-statistical-control with respect to the D statistic. On the other hand, if the null hypothesis for the T^2 and the Q statistics is rejected, the process is out-of-statistical-control. Rejecting the null hypothesis is followed by Step 3, identifying which of the variables are affected by the AOC.

Step 3: Identifying Affected Variables After detecting an AOC, the next step is to determine what has caused this anomalous process behavior. Traditionally, contribution charts and reconstructing process variables and fault conditions along specific directions have been proposed in the literature.[2] For multimodal processes, the data model in Eqs. 1, 2, and 5 can be revisited to develop a reconstruction scheme involving the first m score variables. The extended data model describing a fault condition is of the form

$$\mathbf{z}^*(k) = \mathbf{A}\mathbf{s}^*(k) + \mathbf{e}(k) = \mathbf{A}(\mathbf{s}(k) + \Delta\mathbf{s}) + \mathbf{e}(k) \tag{12}$$

Inspecting the MLPCA model of the n score variables yields

$$\begin{aligned}\mathbf{z}^*(k) &= \mathbf{P}\mathbf{t}(k) + \mathbf{f}(k) = \mathbf{P}(\mathbf{t}(k) + \Delta\mathbf{t}) + \mathbf{f}(k)\\ &= \mathbf{P}_1(\mathbf{t}_1(k) + \Delta\mathbf{t}_1) + \mathbf{P}_2(\mathbf{t}_2(k) + \Delta\mathbf{t}_2) + \mathbf{f}(k).\end{aligned} \tag{13}$$

Here, $\mathbf{P}_1 \in \mathbb{R}^{N\times m}$, $\mathbf{P}_2 \in \mathbb{R}^{N\times(n-m)}$, $\mathbf{t}_1 \in \mathbb{R}^m$, and $\mathbf{t}_2 \in \mathbb{R}^{n-m}$. Fault conditions that manifest themselves in the MLPCA residuals can be diagnosed using contribution charts or variable reconstructions based on the Q statistic, for example, discussed in Ref. 2, and hence, are not discussed here. Given that the loading vectors are mutually orthogonal and of unit length, the effect of the fault condition upon the score variables $\mathbf{t}_1$ and $\mathbf{t}_2$ can be examined individually

$$\mathbf{t}_1(k) + \Delta\mathbf{t}_1 = \mathbf{P}_1^T(\mathbf{z}^*(k) - \mathbf{f}(k)) \cong \mathbf{P}_1^T\mathbf{z}^*(k) \tag{14a}$$

$$\mathbf{t}_2(k) + \Delta\mathbf{t}_2 = \mathbf{P}_2^T(\mathbf{z}^*(k) - \mathbf{f}(k)) \cong \mathbf{P}_2^T\mathbf{z}^*(k) \tag{14b}$$

It follows from Eq. 6 that $\mathbf{t}_1(k) \sim \sum_{i=1}^{M}\pi_i\mathcal{N}\left\{\bar{\mathbf{t}}_1^{(i)}, \mathbf{S}_{tt}^{(i)}\right\}$ and $\mathbf{t}_2(k) \sim \mathcal{N}\{\mathbf{0}, \Lambda_\mathbf{2}\}$. As for the Q statistic, fault diagnosis using contribution charts and variable reconstruction can be obtained for the T^2 statistic. The impact of the fault upon the score variable set $\mathbf{t}_1^*(k)$ can be directly estimated from Eq. 14a

$$\Delta\mathbf{t}_1 \cong E\left\{\mathbf{P}_1^T\mathbf{z}^* - \bar{\mathbf{t}}_1^{(i)}\right\} = \mathbf{P}_1^T E\{\mathbf{z}^*\} - \bar{\mathbf{t}}_1^{(i)}, \tag{15}$$

if the process operates at the ith operating point provided that a sufficiently large number of samples are available to estimate $E\{\mathbf{z}^*\}$. The estimate of $\Delta\mathbf{t}_1$ in Eq. 15 allows constructing conventional contribution charts and variable reconstruction with respect to the D statistic. It is important to note, however, that the samples of the m score variables follow a Gaussian distribution that is described by the mean vector $\bar{\mathbf{t}}_1^{(i)}$ and the covariance matrix $\mathbf{S}_{tt}^{(i)}$, which is generally not a diagonal matrix. It is, therefore, required to determine the fault contribution for the scores by removing correlation as follows

$$\Delta\mathbf{z} = \mathbf{P}_1\left[\mathbf{L}^{(i)}\right]^{-1}\Delta\mathbf{t}_1 \tag{16}$$

where $\mathbf{S}_{t_1t_1}^{(i)} = \mathbf{L}^{(i)}[\mathbf{L}^{(i)}]^T$ is the Cholesky decomposition of $\mathbf{S}_{t_1t_1}^{(i)}$.

Alternative data structures

So far, we have assumed that the proposed method relies on the existing data described in Eq. 1. It is practically possible, however, that this data structure is too restrictive. More precisely, the following two scenarios are possible: (1) $n = m = N$ and (2) $n = N > m$. In the former case, the number of source signals is identical to the number of recorded variables and the number of components required describing between-cluster variation. Hence, there are no components that relate to within-cluster variation, which is the situation assumed in Ref. 9. In the latter case, the number of source signals is equal to the number of recorded variables with the number of components describing between-cluster variation being smaller than that. In other words, there are $N - m$ components that refer to within-cluster variation.

In the case $n = m = N$, the benefit of the proposed approach lies in the component-wise extraction of the individual modes, which follows from the preceding discussion. Compared to the work in Ref. 9, this circumvents the use of the EM algorithm to identify a GMM model. In the latter case, the proposed work in this article can extract the m and $N - m$ components to describe between- and within-cluster variation, respectively. It is important to note, however, that both alternative data structures require the application of PCA before the application of the component-wise analysis. It is also important to note that the data structure in Eq. 1 also describes both of these cases.

Simulation Example

The first comparison of the proposed multimodal monitoring scheme with existing work is based on the simulation of a CSTR process. This is a jacketed CSTR system to perform the reaction $A \rightarrow B$ for the parameters listed in Table 1.

Table 2 lists the 10 recorded variables and the van Heerden diagram, shown in Figure A1, highlights that this system has a total of three operating modes. The second operating mode is unstable and requires regulatory controllers for reactor volume and reactor temperature. Both controllers have proportional action, where the volume and temperature are controlled by the reactor outlet flow and the jacket flow, respectively. The controller gains for manipulating the reactor outlet flow and the jacket flow are $P_V = -10$ and $P_T = -1$, respectively. A detailed mechanistic model of the CSTR simulator is given in the appendix.

Table 1. Parameters for CSTR Simulation

Parameters	Description	Values	Units
ρ	Density of the stream liquid	50	lb_m/ft^3
ρ_j	Density of the cooling liquid	62.5	lb_m/ft^3
Cp	Heat capacity of the stream liquid	0.75	BTU/lb_m R
Cp_j	Heat capacity of the cooling liquid	1	BTU/lb_m R
V_j	Jacket volume	3.85	ft^3
U	Overall heat transfer in the jacket	150	BTU/h ft^2 R
A	Heat transfer area	250	ft^2
k_0	Pre-exponential factor	7.08×10^{10}	1/h
E	Activation energy	30000	BTU/mol
R	Universal gas constant	1.99	BTU/mol R
ΔH	Reaction enthalpy	−30000	BTU/mol

From the list of variables in Table 2, the presented analysis did not include the jacket flow and the reactor outlet flow rate, as they are manipulated variables which are driven by proportional controllers. To account for sensor and process noise, the remaining eight variables were augmented by zero mean independently distributed Gaussian sequences of variance 0.3 (fresh feed flow), 0.001 (fresh feed reactant concentration), 0.25 (fresh feed temperature), 0.2 (inlet jacket temperature), 0.5 (reactor volume), 0.001 (reactant concentration), and 0.25 (reactor temperature and 0.2 (jacket temperature).

Identifying a multimodal monitoring model

To identify a multimodal monitoring model and to test the model for fault detection, two data sets were recorded. Both sets covered 48 h of data, recorded at a frequency of one sample every 150 s (1152 samples). The first data set covers normal process behavior. Figure 2a outlines that the application of the equivalence of eigenvalues test[28] suggested the retention of $n = 4$ principal components.

This was expected, given that the fresh feed flow, the fresh feed reactant concentration, the fresh feed temperature, and the inlet jacket temperature are measured disturbances for the system. The MLPCA algorithm produced the following estimate for error covariance matrix

$$\text{diag}\left\{\hat{\mathbf{S}}_{ee}\right\} = (0.3723\ 0.0006\ 0.2380\ 0.1958\ 0.4998\ 0.0011\ 0.2797\ 0.2135)^T. \quad (16)$$

which is close to the actual diagonal parameters in $\mathbf{S}_{ee}$. In Eq. 16, the operator diag $\{\cdot\}$ returns the diagonal elements of a matrix. It should be noted that the nondiagonal elements of $\mathbf{S}_{ee}$ are equal to zero.

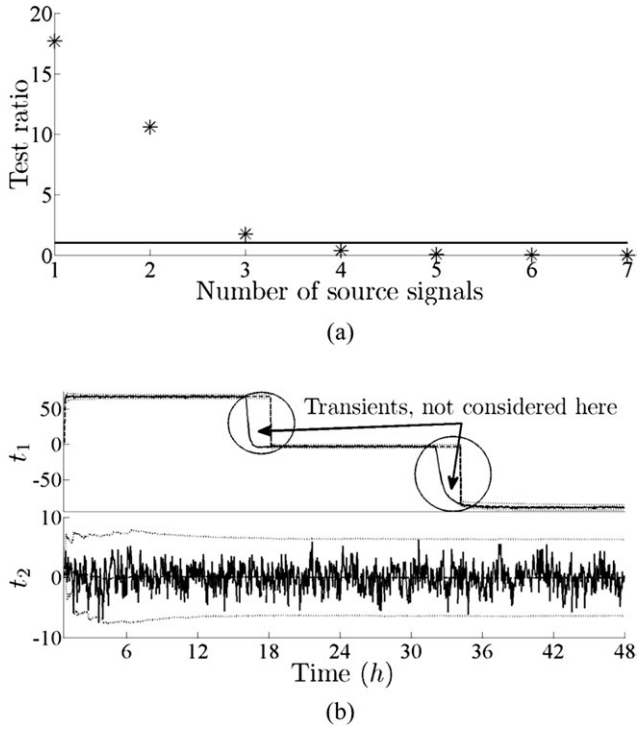

Figure 2. Result of applying equivalence of eigenvalues test (a) and plots of first two score variables (b) for CSTR simulation example.

After establishing a MLPCA model, the next step involved the identification of a multimodal model on the basis of the four retained principal components. Figure 2b highlights that the first component describes the $M = 3$ modes, so that $m = 1$. In contrast, the remaining $n - m = 3$ components do not describe the different operating modes. The subsection now contrast the performance of the proposed multimodal monitoring scheme with standard MLPCA-based method, discussed in the preliminaries of this article, and the multimodal GMM approach advocated in Ref. 9.

Diagnosis of a catalyst degradation

The second data set describes catalyst deactivation by manipulating the pre-exponential constant of the Arrhenius equation in Eq. A5. Figure 3a shows that the catalyst deactivation started from 16 h into the data set and resulted in a gradual reduction of the constant by 50% over a period of 16 h. During the recording period, the process operated in operation Mode 2, which Figure 3d indicates. Figure 3b presents the conventional

Table 2. Recorded Variable Set for CSTR Simulation

Variables	Description	Values	St. dev.		Units
F_o	Fresh feed flow	40	2		ft^3/h
C_{Ao}	Fresh feed reactant concentration	0.5	0.1		mol/ft^3
T_o	Fresh feed temperature	530	4		R
T_{jo}	Inlet jacket temperature	530	3		R
F_j	Jacket flow	49.9	–		ft^3/h
F	Reactor outlet flow	40	–		ft^3/h
V	Reactor volume	48	–		ft^3
Variables	Description	First Mode	Second Mode	Third Mode	Units
C_A	Reactant concentration	0.4739	0.2451	0.0591	mol/ft^3
T	Reactor temperature	537.16	599.99	651.06	R
T_j	Jacket temperature	536.62	594.63	641.79	R

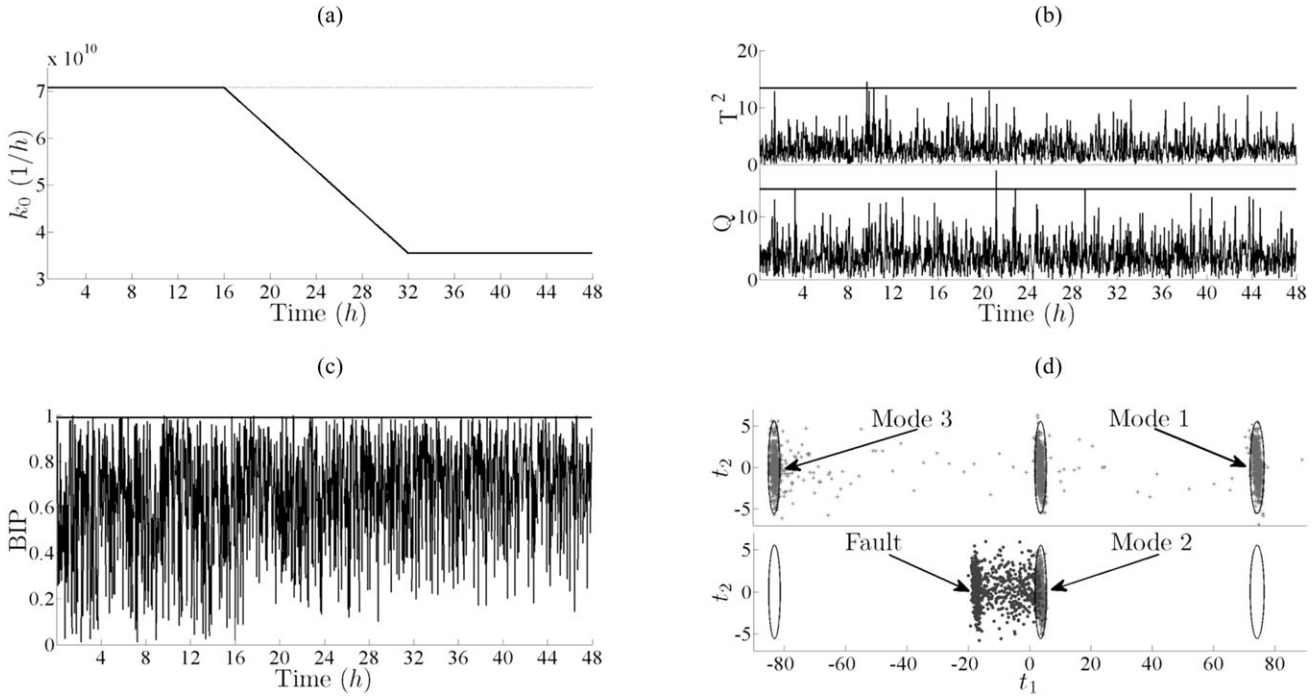

Figure 3. Results of detecting catalyst degeneration techniques: (a) change in reaction constant, (b) application of MLPCA, (c) application of GMM, and (d) effect of fault condition in score space.

Hotelling's T^2 and Q statistics, based on the MLPCA model, and highlights that the fault condition could not be detected. The same picture emerged for the BIP statistic (Figure 3c) of the GMM model, defined by Eqs. 28–30 in Ref. 9, which is also insensitive to this abnormal event.

More precisely, the BIP statistic, which uses an implicit distance rule, suggested that the process is behaving normally, as the fault has no major impact on the individual process variables. In contrast, the proposed approach, which relies benchmarking the current operating condition against a specific operating mode, was able to detect the catalyst deactivation based on the first score variable, which follows from Figure 3d.

As a result, the D statistic of the proposed monitoring scheme, monitoring the between-cluster variation of the second operating model, detected this event shortly after the catalyst deactivation was initiated, which Figure 4a outlines. It follows from Figure 3a that the initial phase of the deactivation was a gradual deterioration, which yields an average run length of around 1 h.

This can also be noted from Figure 3d, as a number of scatter points are still within the control ellipse describing the second operating mode. The Hotelling's T^2 statistic of the proposed scheme, monitoring the within-cluster variation was insensitive to this event. This result confirms the insensitivity of the conventional Hotelling's T^2 statistic in Figure 3b. The residual Q statistics for the conventional MLPCA approach and the proposed monitoring scheme are identical and also insensitive to this fault condition. This, in turn, implies that the fault condition only affects the orthogonal projections of the samples onto the MLPCA model subspace.

In summary, the analysis of this simulation example showed that the proposed multimodal monitoring scheme is more sensitive than conventional PCA and the recently proposed method in Ref. 9. The main advantage is the distinction between within- and between-cluster variations based on the data structure in Eq. 5.

Computational efficiency of proposed approach

As this example showed, applying the component-wise identification of operation modes reduced the task of identifying a GMM involving $N = 10$ process variables down to $m = 1$ score variable describing between-cluster variation. It

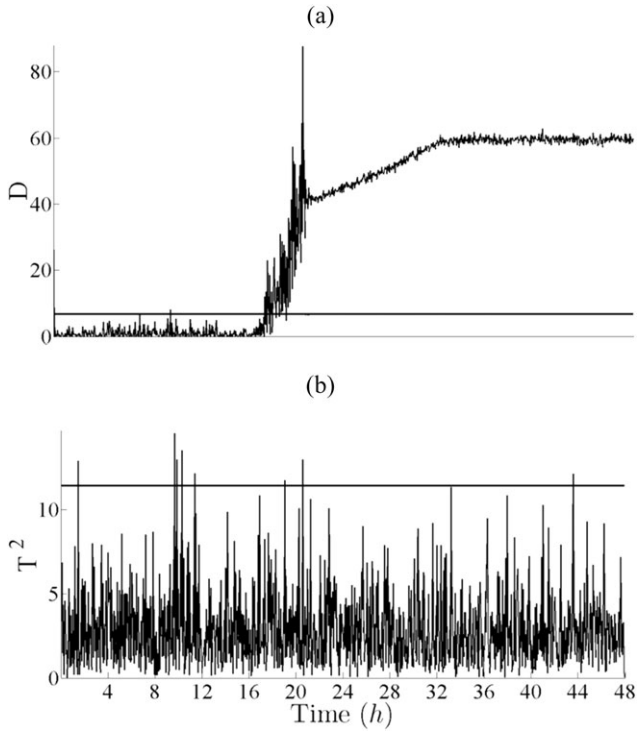

Figure 4. Results of applying proposed approach: (a) *D* statistic monitoring individual cluster and (b) Hotelling's T^2 statistic monitoring within-cluster variation.

is also interesting to examine whether the application of the classification technique, proposed in Figure 1, is computationally more efficient than the EM algorithm.[28] To do so, the article contrasts the performance of both approaches for determining a GMM for the $m = 1$ score variable describing the multimodal behavior of this process.

Given that the EM algorithm is an iterative approach and depends on the starting point, a total of 10 iterations with randomized initial conditions were carried out and the average time was estimated. The average time was also determined for the proposed approach by running the algorithm in Figure 1 a total of 10 times. This produced an average time of 1.77 s for the EM algorithm and 0.19 s for the algorithm in Figure 1. Thus, the algorithm in Figure 1 requires 89% less time to compute the GMM compared to the EM algorithm.

It is important to note that the computational burden for the EM algorithm is expected to increases significantly with the number of components representing between-cluster variation. In contrast, the classification algorithm in Figure 1 only yields a linear increase with the number of components.

Application to a Furnace Process

This process is an intermediate furnace of a catalyst reforming process (Powerforming) used to regulate the temperature of the reactants mixture between consecutive reactors. There are a total of eight temperature sensors for which a reference data set covering approximately 8 days at a sampling rate of 30 s (~23,000 samples). Figure 5a shows that this process has two distinct operating modes.

The objective of this application study is to construct a multimodal monitoring model for the proposed monitoring scheme. As the eight temperature sensors are identical, it is assumed here that the error variance for each temperature variable is identical (isotropic error covariance matrix). This can be confirmed by the eigendecomposition of the data covariance matrix, which produces the following eigenvalues

$$\text{diag}\{\Lambda\} = (\,7.98 \quad 0.00 \quad 0.00 \quad 0.00 \quad 0.00 \quad 0.00 \quad 0.00 \quad 0.00\,)^T. \quad (17)$$

As before, the operator diag $\{\Lambda\}$ returns a vector storing the diagonal elements of the matrix of eigenvalues Λ. The first eigenvector (principal direction) of the data covariance matrix stores, as expected, almost identical elements

$$\mathbf{p}_1^T = (\,0.35 \quad 0.35 \quad 0.35 \quad 0.35 \quad 0.35 \quad 0.35 \quad 0.35 \quad 0.35\,) \quad (18)$$

The application of PCA, therefore, outlines that the first components represents common cause variation, which describes between-cluster variation, that is $m = 1$. On the other hand, the remaining seven components span the residual subspace, such that $n = m = 1$. The next step involved the identification of the GMM parameters for the first principal component. The adaptive approach, detailed in Figure 1, suggests that $M = 2$ and the estimates for π_i, $\bar{t}^{(i,1)}$ and $\sigma^{(i,1)}$ are

$$\begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} 0.44 \\ 0.56 \end{pmatrix} \quad (19a)$$

$$\begin{pmatrix} t^{(1,1)} \\ t^{(2,1)} \end{pmatrix} = \begin{pmatrix} -3.1 \\ 2.4 \end{pmatrix} \quad (19b)$$

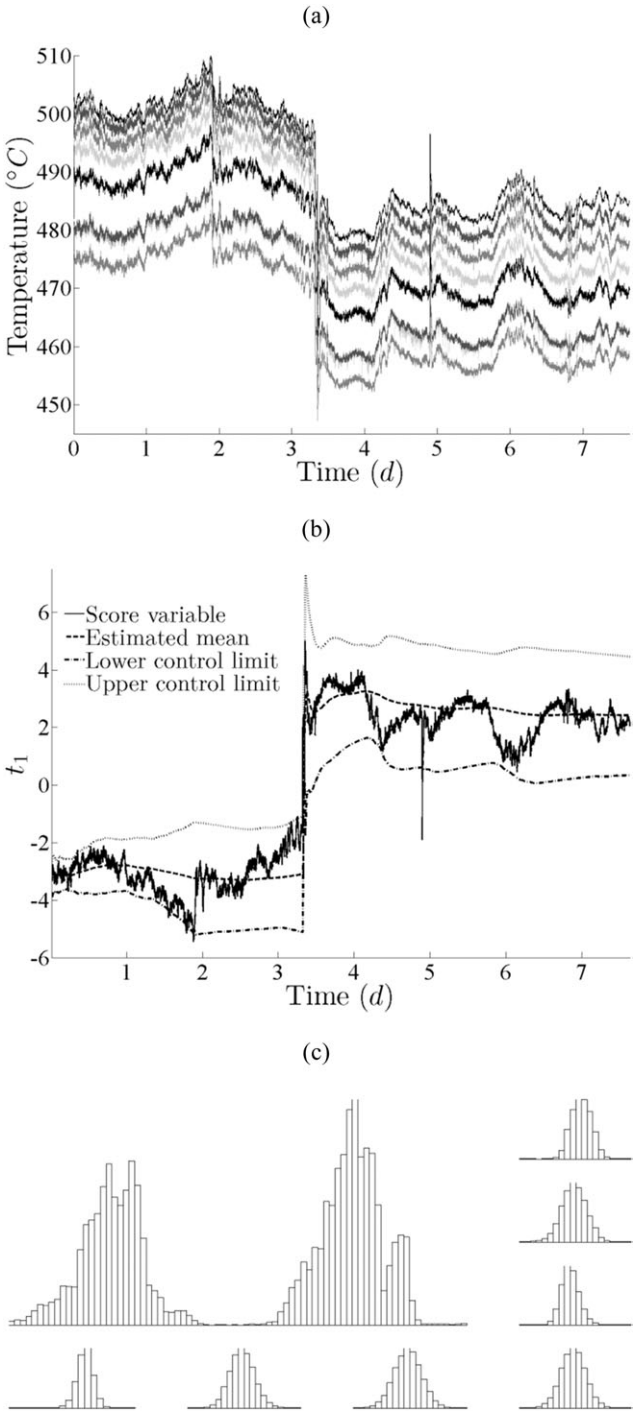

Figure 5. Results of applying proposed multimodal monitoring scheme: (a) time-based plots for each of the eight temperature variables, (b) clustering approach to determine the two operating modes, and (c) histograms for eight score variables indicating that only the first one describes the two modes.

$$\begin{pmatrix} \sigma^{(1,1)} \\ \sigma^{(2,1)} \end{pmatrix} = \begin{pmatrix} 0.50 \\ 0.48 \end{pmatrix} \quad (19c)$$

Figure 5b shows the Shewhart chart of the first score variable including the adaptive control limits. The benefit of applying the proposed multimodal monitoring scheme becomes clear when plotting the histograms for the eight principal

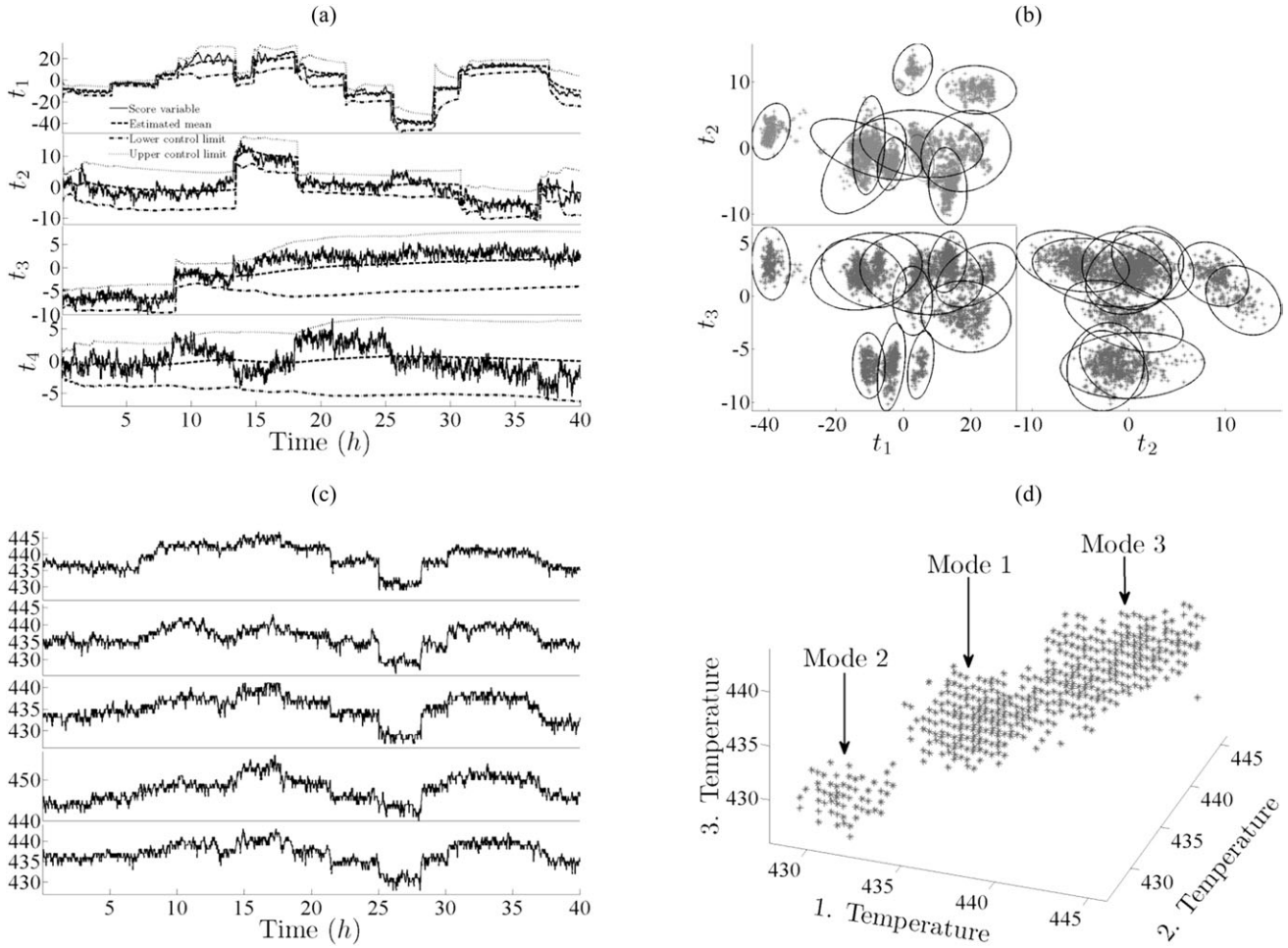

Figure 6. Results of applying monitoring approaches to recorded data of chemical reaction process: (a) application of proposed clustering method, (b) scatter diagrams of first three score variables showing different clusters, detecting process fault using (c) conventional MLPCA approach, and (d) GMM clustering approach.

components, which Figure 5c depict. The larger histogram is related to the first component, whereas the surrounding seven histograms represent the remaining seven components. This is in sharp contrast to the approach introduced in Ref. 9, which is multimodal but with respect to the original variable set containing eight temperature variables and not one principal component.

Application to a Chemical Reaction Process

This process is comprised of several operating units and produces two solvent chemicals. The reactions take place in a number of parallel operating fluidized bed reactors that receive a total of five input streams (reactants) from supply storage, upstream units, and plant recycles. Each of these reactors consists of a large shell and 35 vertically oriented tubes in which the complex reactions take place supported by fluidized catalyst. As the reactions are exothermic, oil circulates around the tubes as a coolant. The temperature in each of the 35 tubes is measured by a thermocouple located at the bottom of each tube. A more detailed discussion of this process is available in Refs. 2 and 26.

The operation of this process can be affected by a number of unmeasured disturbances introduced to the vaporizer unit and the coolant. Changes in the catalyst fluidization also affect the chemical reaction and result in an increase or decrease in the measured temperature. From this process, two data sets were analyzed; a reference set of around 40 h of data ($\sim$2400 samples) and a set of about 50 h ($\sim$3000 samples) describing a drop in steam pressure, which affects the vaporizer. The process has a number of distinct operating conditions (modes) and the aim of this section is to compare the proposed multimodal monitoring scheme with conventional MLPCA monitoring and the method proposed in Ref. 9.

Identifying a multimodal monitoring model

The application of MLPCA suggested the number of source signals to be $n = 20$[28] and Table 6.9 in Ref. 3 lists the diagonal elements of the estimated error variance matrix. The application of the adaptive approach to each of the $n = 20$ components suggested that the first $m = 3$ components describe $M = 11$ distinct operating conditions, which Figures 6a, b show. Hence, the first $m = 3$ components describe between-cluster variation and the remaining $n - m = 17$ components represent the within-cluster variation. Moreover, the model and residual subspaces are of dimension 20 and 15, respectively.

Although there appears to be some significant overlap among the individual control ellipses for the two-dimensional (2-D) representations, the individual clusters only have some minor overlap, when examined in a 3-D plot, which is, however, difficult to show. Figure 8b gives a

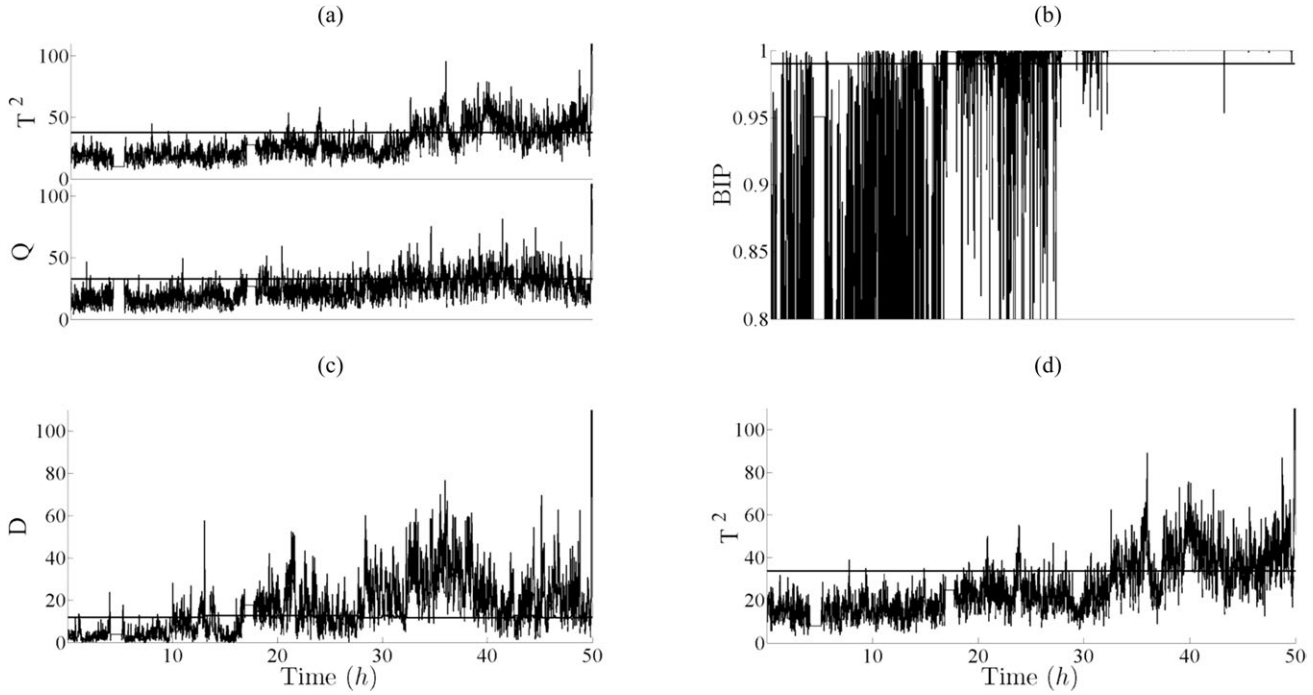

Figure 7. Results of applying proposed approach: (a) *D* statistic monitoring individual cluster, (b) Hotelling's T^2 statistic monitoring within-cluster variation, (c) identified operating modes, and (d) residual-based *Q* statistic.

graphical account of the first three modes (clusters) and confirms that there is little overlap between them. In contrast, application of the GMM technique based on the set of 35 tube temperatures identified a total of only three operating modes only. Figure 6c depicts the recorded signals of the first five temperatures and Figure 6d indicate the presence of only three modes (clusters) when based on the first three variables.

It is important to note, however, that Figures 6c, d can only be seen as an illustration to outline that the application of the GMM technique to $N = 35$ temperature variables may not reveal the correct underlying number of modes. In contrast, Figure 6a presents a clear picture to support that there is a larger number of different operating models modes, which particularly the first score variable shows.

Diagnosis of a drop in steam pressure

Steam provides the required enthalpy for vaporizing some of the feed streams before entering the reactor. Although the temperature of the steam remained constant, the drop in steam pressure affected the enthalpy balance within the vaporizer. Consequently, less heat was transferred, which, in turn, lowered the enthalpy with which the feed streams entered the reactor. This had also some effect on the reaction conditions and hence, the tube temperatures.

Figure 7a shows the results of applying a conventional MLPCA-based monitoring model. The residual Q statistic is more sensitive and detected this event after about 17 h into the data set. The Hotelling's T^2 statistic detects the effect of the drop in stem pressure after 30 h although some sporadic violations can be noticed from around 20 h into the data set. Figure 7b summarizes the results of applying the GMM-based approach, discussed in Ref. 10. Significant violations of the control limit for the BIP statistic (upper plot) arose between 1 and 4 h, between 9 and 14 h, and after 17 h into the data set. Figures 7c, d show the performance of the D and T^2 statistic of the proposed multimodal scheme, respectively. The D statistic detected abnormal events between 10 and 13 h and from around 16 h into the data set; whist the T^2 statistic shows violations from around 20 h into the data set.

As the drop in steam pressure did not yield a considerable upset in the performance of the reactor and the chemical reactions in particular, the recorded temperature variables did not show a significant departure from normal operating patterns. Moreover, it was later determined that the drop in steam pressure arose between 14 and 15 h into this second data set. Any event that was detected by the GMM technique in Ref. 10 and proposed method can, therefore, not be linked to this process fault. What can be stated by this comparison, however, is that the proposed method is more sensitive in detecting this event, that is, 16 h compared to 17 h for the technique in Ref. 9 and the application of a conventional monitoring model.

Figure 8 analyzes the abnormal events detected by both multimodal techniques in more detail. The plot in Figure 8a shows a 3-D scatter diagram based on the first three temperature variables covering the first 5 h of data. The control ellipses were obtained from the original 35-D ellipsoid. Initially, the technique in Ref. 10 detected correctly the first operating mode but a number of samples were identified to be outside any of the three modes, which the plot in Figure 8c confirms. Again, the plot in Figure 8a can only be used as an illustration, as the GMM model relies on all 35 temperature variables. However, the points marked by "•" in this plot were not earmarked as belonging to Mode 1. Mode 0 is referred to in this analysis as an unknown mode that was not present in the reference data.

Figure 8b shows the scatter points of the first three score variables for the first 14 h. After about 5 h, the operational model changed from the initial Mode 1 to Mode 2 and during the period between 10 and 13 h a transient from

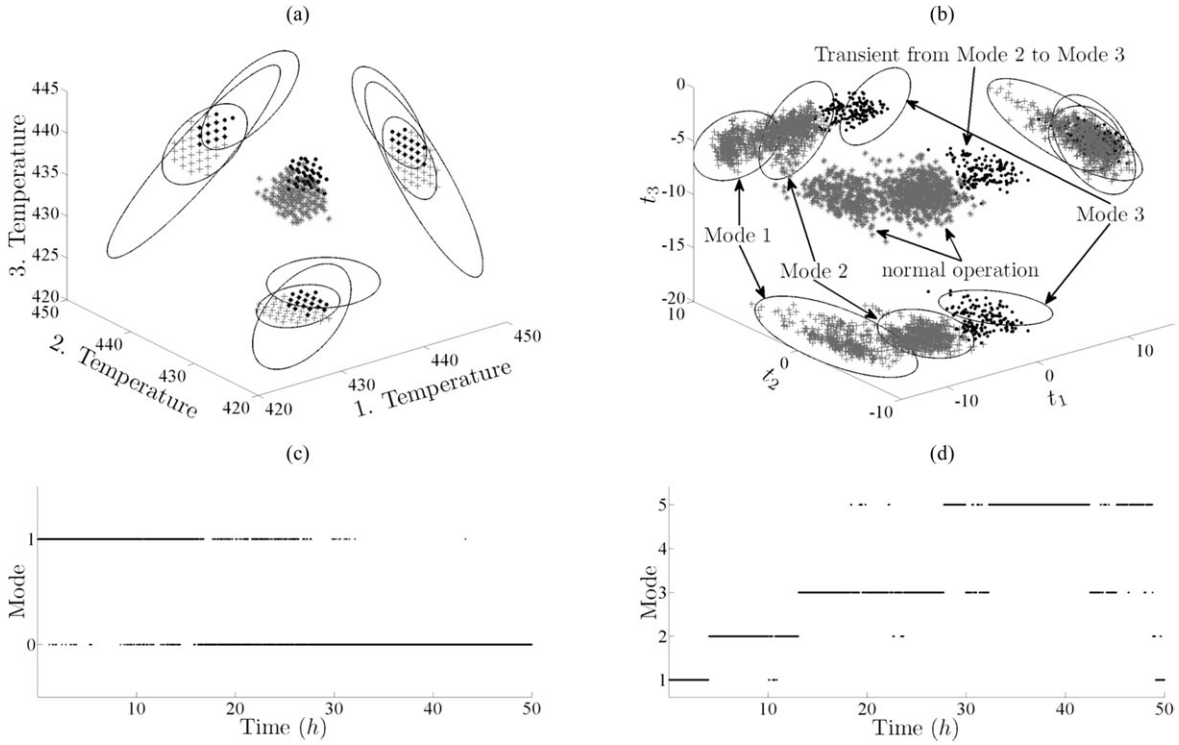

Figure 8. Scatter plots showing first three modes and transients between them for first 18 h of data describing fault condition.

Modes 2 and 3 arose. This transient was the main reason behind the violation of the *D* statistic. During this transient, the operating mode was occasionally detected to be Model 1. This highlights the limitation of the proposed work that is not designed to describe transient behavior. To deal with transient changes, however, the work discussed in Ref. 29 can be used to augmenting the proposed multimodal technique, which is beyond the scope of this article.

Concluding Summary

This article has described and analyzed existing work on monitoring complex multimodal processes in the chemical industry. The analysis has shown that existing work can be categorized into: (1) mixture of component or Gaussian models, (2) the assumption of common structures for operating modes, (3) hybrid approaches, and (4) alternative approaches. The analysis also highlighted that the following issues have not been adequately presented in the literature: (1) how to deal with a general assumption about the error covariance matrix in a PCA context, (2) how to take advantage of a reduced dimensional data representation in a multimodal context, (3) how to determine operating modes automatically based on the PCA score variables, and (4) how to separate variation that the operating modes commonly share from variation that is specific to individual clusters.

To address these issues, this article proposes the use of a common noncausal data structure. The source variables of the data structure describe between-cluster variation that is specific for the individual modes but also within-cluster variation which is common to each of the clusters. Further assumption is that error vector for each of the modes has a common error covariance matrix and that the error variables are statistically independent of the source variables. After establishing a MLPCA model, the score variables are individually analyzed to reveal the different operating modes and to determine whether they are describing between- or within-cluster variation. Once the number of score variables describing between-cluster variation is obtained, a GMM can be identified. The article has highlighted that this clustering approach overcomes the problems associated with the EM algorithm in determining GMMs, which emphasizes the practical value of the proposed approach.

The proposed multimodal monitoring model requires a total of three nonnegative squared statistics. The first one monitors between-cluster variation relative to the current operating mode that is automatically detected. The second one monitors the within-cluster variation, whereas the remaining one is a residual-based statistic that is based on the PCA model residuals. The article shows the benefits of relying on a reduced dimensional data representation instead of the applying GMM model for the original variable space through the application of a simulated CSTR process and two application studies that involve the analysis of recorded process data from a furnace process and a chemical reaction process. For the simulated process, the application of the proposed multimodal scheme has shown to be more sensitive than conventional work in detecting simulated catalyst deterioration. Moreover, the proposed scheme has been able to determine a significantly larger number of modes than the application of a GMM model to a total of 35 recorded temperature variables. This translated in a more accurate and sensitive process monitoring. The applications studies have also outlined, however, that transients between different operating models can be detected incorrectly as fault conditions. This can be overcome by blending existing work incorporating the presence of transient changes into the proposed monitoring scheme.

Future work on this topic could include its applications to systems that present nonlinear relationships between the recorded process variables, implying that the clusters do not lie on a linear model subspace but a nonlinear model surface.

This could be addressed by using an autoassociative neural network or kernel PCA to identify a nonlinear PCA model.

Literature Cited

1. Nimmo I. Adequately address abnormal situation management. *Chem Eng Progress*. 1995;91:36–45.
2. Kruger U, Xie L. *Statistical Monitoring of Complex Multivariate Processes*. Chichester: Wiley, 2012.
3. MacGregor JF, Marlin TE, Kresta J, Skagerberg B. *Multivariate statistical methods in process analysis and control*. In: *AIChE Symposium Proceedings of the 4th International Conference on Chemical Process Control*, New York, USA, 1991: P-67:79–99.
4. Wise BM, Gallagher NB. The process chemometrics approach to process monitoring and fault detection. *J Process Control*. 1996; 6:329–348.
5. Wang X, Kruger U, Lennox B. Recursive partial least squares algorithms for monitoring complex industrial processes. *Control Eng Pract*. 2003;11:613–632.
6. Gertler J, Li WH, Huang YB, McAvoy T. Isolation enhanced principal component analysis. *AIChE J*. 1999;45:323–334.
7. MacGregor JF, Yu H, Muñoz SG, Flores-Cerrillo J. Data-based latent variable methods for process analysis, monitoring and control. *Comput Chem Eng*. 2005;29:1217–1223.
8. AlGhazzawi A, Lennox B. Monitoring a complex refining process using multivariate statistics. *Control Eng Pract*. 2008;16:294–307.
9. Yu J, Qin J. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J*. 2008;54:1811–1829.
10. Lane S, Martin EB, Kooijmans R, Morris AJ. Performance monitoring of a multiproduct semi-batch process. *J Process Control*. 2001;11:1–11.
11. Chen J, Liu J. Mixture principal component analysis models for process monitoring. *Ind Eng Chem Res*. 1999;38:1478–1488.
12. Zhao SJ, Zhang J, Xu YM. Performance monitoring of processes with multiple operating modes through multiple PLS models. *J Process Control*. 2006;16:763–772.
13. Ng YS, Srinivasan R. An adjoined multimodel approach for monitoring batch and transient operations. *Comput Chem Eng*. 2009; 33:887–902.
14. Natarajana S, Srinivasan R. Multimodel based process condition monitoring of offshore oil and gas production process. *Chem Eng Res Design*. 2010;88:572–591.
15. Tipping ME, Bishop CM. Mixtures of probabilistic principal component analyses. *Neural Comput*. 1999;11:443–482.
16. Choi SW, Martin EB, Morris AJ, Lee IB. Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture. *Ind Eng Chem Res*. 2005;44:2316–2327.
17. Ge Z, Song Z. Maximum-likelihood mixture factor analysis model and its application for process monitoring. *Chemom Intell Lab Syst*. 2010;102:53–61.
18. Maestri M, Farall A, Groisman P, Cassanello M, Horowitz G. A robust clustering method for detection of abnormal situations in a process with multiple steady-state operation modes. *Comput Chem Eng*. 2010;34:223–231.
19. Choi SW, Park JH, Lee IB. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Comput Chem Eng*. 2004;28:1377–1387.
20. Chen T, Zhang J. On-line multivariate statistical monitoring of batch processes using Gaussian mixture model. *Comput Chem Eng*. 2010;34:500–507.
21. Yu J. Hidden Markov models combining local and global information for nonlinear and multimodal process monitoring. *J Process Control*. 2010;20:344–359.
22. Ge Z, Yang C, Song Z, Wang H. Robust online monitoring for multimode processes based on nonlinear external analysis. *Ind Eng Chem Res*. 2008;47:4775–4783.
23. Ge Z, Song Z. Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Control Eng Pract*. 2008;16:1427–1437.
24. Ge Z, Song Z. Multimode process monitoring based on Bayesian method. *J Chemometrics*. 2009;23:636–650.
25. Camacho J, Pico J, Ferrer A. data understanding with PCA: structural and variance information plots. *Chemom Intell Lab Syst*. 2010;100:48–56.
26. Feital TS, Kruger U, Xie L, Schubert U, Lima EL, Pinto JC. A unified statistical framework for monitoring multivariate systems with unknown source and error signals. *Chemom Intell Lab Syst*. 2010;104:223–232.
27. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst*. 1987;2:37–52.
28. Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. *IEEE Trans Pat Anal Mach Intell*. 2002;24:381–396.
29. Kourti T. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J Chemometrics*. 2003;17:93–109.
30. Wang X, Kruger U, Lennox B. Process monitoring approach using fast moving window PCA. *Ind Eng Chem Res*. 2005;44:5691–5702.
31. Jackson JE, Mudholkar GS. Control procedures for residuals associated with principal component analysis. *Technometrics*. 1979;21: 341–349.
32. Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. *Technometrics*. 1995;37:41–59.

Appendix A: Jacketed CSTR Simulator

The mathematical model for the CSTR process is as follows

$$\frac{dV}{dt} = F_i - F \tag{A1}$$

$$\frac{dVC_A}{dt} = F_i C_{A_i} - FC_A - VrC_A \tag{A2}$$

$$\frac{dVT}{dt} = F_i T_i - FT - \frac{\Delta H}{\rho c_p} VrC_A - \frac{UA}{\rho c_p}(T - T_j) \tag{A3}$$

$$\frac{dT_j}{dt} = \frac{F_j(T_{j_i} - T_j)}{V_j} + \frac{UA}{\rho_j c_{p_j} V_j}(T - T_j) \tag{A4}$$

$$r = r_0 e^{-\frac{E}{RT}} \tag{A5}$$

$$F = F_{\text{set}} + P_V(V_{\text{set}} - V) \tag{A6}$$

$$F_j = F_{j_{\text{set}}} + P_T(T_{\text{set}} - T) \tag{A7}$$

In the above equations, r is the reaction rate, the subscript "set" refers to a setpoint, and the proportional controller gains P_V and P_T are -10 and -1, respectively. Figure A1 shows the enthalpy balance for the reactor, which confirms that the CSTR reactor has three distinct operating modes.

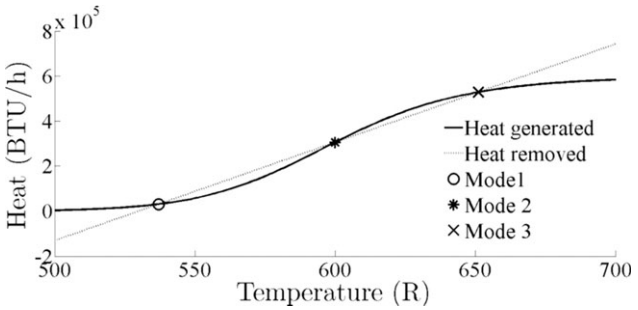

Figure A1. van Heerden diagram for CSTR process.

Manuscript received Jun. 26, 2012, and revision received Sept. 4, 2012.