IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 8, 2017, accepted August 28, 2017, date of publication September 6, 2017, date of current version September 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2749498

Data-Driven Optimized Distributed Dynamic PCA for Efficient Monitoring of Large-Scale Dynamic Processes

YANG WANG[1,2], QINGCHAO JIANG[3], AND JINGQI FU[2]
[1]School of Electric Engineering, Shanghai Dianji University, Shanghai 200240, China
[2]School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China
[3]Key Laboratory of Advanced Control and Optimization for Chemical Processes of Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

Corresponding author: Qingchao Jiang (qchjiang@ecust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603138, in part by Fundamental Research Funds for the central universities under Grant 222201717006 and Grant 222201714027, in part by the Young Teacher Study Abroad Program of Shanghai under Grant A1-0217-16-003-01, and in part by the Shanghai Pujiang Program under Grant 17PJD009.

ABSTRACT Dynamic principal component analysis (DPCA) is generally employed in monitoring dynamic processes and typically incorporates all measured variables. However, for a large-scale process, the inclusion of variables without fault-relevant information may cause redundancy and degrade monitoring performance. In this paper, the influence of variable and time-lagged variable selection on the DPCA monitoring performance is analyzed. Then, a fault-relevant performance-driven distributed monitoring scheme is proposed to achieve efficient fault detection and diagnosis. First, performance-driven process decomposition is performed, and the optimal subset of variables and time-lagged variables for each fault are selected through a stochastic optimization algorithm. Second, local DPCA models are established to characterize the process dynamics and generate fault signature evidence. Finally, a Bayesian diagnosis system with the most efficient evidence sources is established to identify the process status. Case studies on a numerical example and the Tennessee Eastman benchmark process demonstrate the efficiency of the proposed monitoring scheme.

INDEX TERMS Distributed monitoring, large-scale dynamic process, dynamic principal component analysis, Bayesian fault diagnosis.

I. INTRODUCTION

Multivariate statistical process monitoring (MSPM) is gaining increasing attention because of the increasing demands in plant safety and product quality [1]–[5]. Among the MSPM methods, principal component analysis (PCA) generally serves as one of the most fundamental techniques. A static PCA efficiently handles the cross-correlation among variables. However, practical processes can be characterized by dynamics, and some degree of auto-correlation exists among variables. To address the dynamic behaviors, several dynamic methods using the ''time lag shift'' strategy such as dynamic PCA (DPCA) has been developed [6]–[9]. Efficiency has been demonstrated by numerous applications, but direct application of these methods on a large-scale dynamic process is not appropriate because of the large number of measured variables and the complex correlations (cross-correlation and auto-correlation).

To address the monitoring issues for large-scale processes, multiblock or distributed methods are generally employed, in which the process decomposition is one of the key steps [10]–[14]. Recently, several data-driven distributed monitoring schemes, which decompose the process relying totally on data analysis, have been developed [15]–[18]. However, without considering the fault information, optimal monitoring performance cannot be guaranteed. Taking fault information into account, the performance-driven distributed monitoring scheme has been developed [19]. However, the influence of variable and time-lagged variable selection on the DPCA monitoring performance has not yet been discussed.

Another issue in monitoring a large-scale process is fault diagnosis, which aims to assign the current sample to the most related historical fault class [1], [20]. As an efficient decision-making technique, Bayesian method is superior in

fault diagnosis, and numerous studies on Bayesian fault diagnosis have been conducted [21]–[23]. Recently, an optimal Bayesian diagnosis system for PCA-based process monitoring was developed [20]. However, inclusion of the Bayesian diagnosis system into the distributed monitoring of a large-scale dynamic process has not been performed.

The main contributions of the current work can be summarized as follows. (i) The influence of process decomposition on the DPCA monitoring performance is theoretically analyzed, which will enhance the dynamic process monitoring theory basics. (ii) A performance-driven process decomposition method, which selects the optimal subset of variables and time-lagged variables to achieve the best possible monitoring performance, is proposed. (iii) An efficient Bayesian diagnosis system with optimal evidence sources is established, which identifies the process status of a dynamic process. The remainder of the paper is structured as follows. Section 2 reviews the basics of DPCA monitoring and Bayesian diagnosis. The influence of process decomposition on the DPCA monitoring performance is also analyzed. Section 3 presents in detail the performance-driven distributed DPCA monitoring scheme. Section 4 provides application results on a numerical example and the Tennessee Eastman (TE) benchmark process. Finally, Sections 5 draws the conclusions.

II. PRELIMINARIES

A. DPCA PROCESS MONITORING

PCA establishes a correlation-concerned model for process monitoring. Given a normalized data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m] \in R^{N\times m}$ with N denoting the number of observations and m denoting the number of measured variables, the PCA transformation matrix $\boldsymbol{U}$ can be obtained by performing eigenvalue decomposition on the covariance matrix $\boldsymbol{\Sigma}$. Given a new sample $\boldsymbol{x} \in R^{m\times 1}$, the i-th principal component (PC) score t_i can be calculated as follows [20]:

$$t_i = \frac{\boldsymbol{x}^T \boldsymbol{p}_i}{\sqrt{\lambda_i}}, \tag{1}$$

where λ_i is the i-th eigenvalue corresponding to the i-th PC, and $\boldsymbol{p}_i \in R^{m\times 1}$ is the i-th eigenvector in the transformation matrix $\boldsymbol{U}$. Two statistics named T^2 and Q are constructed for fault detection as follows [1]:

$$T^2 = \boldsymbol{t}^T\boldsymbol{t} = \sum_{i=1}^{r} t_i^2 \le T_{\lim}^2, \tag{2}$$

$$Q = \left(\boldsymbol{x}-\hat{\boldsymbol{x}}\right)^T\left(\boldsymbol{x}-\hat{\boldsymbol{x}}\right) = \boldsymbol{x}^T\left(\boldsymbol{I}-\boldsymbol{U}_r\boldsymbol{U}_r^T\right)\boldsymbol{x} \le \boldsymbol{Q}_{\lim}, \tag{3}$$

where r is the number of retained PCs, $T_{\lim}^2$ denotes the confidence limit of T^2, $\hat{\boldsymbol{x}} = U_r\boldsymbol{t}$ is the recovered data, $\boldsymbol{U}_r$ is the loading matrix with r retained projecting vectors, and $Q_{\lim}$ is the confidence limit of the Q statistic.

The above equation represents the static PCA monitoring model, which efficiently characterizes the cross-correlation among variables. For a dynamic process, auto-correlation also exists and the current values of variables depend on the past values. The auto-correlation or at least the relationship between the current sample and previous sample should be captured. DPCA takes auto-correlation into account by using the "time lag shift" method [7], [9]. In DPCA, the training data set becomes $\boldsymbol{Z} = [\boldsymbol{X}(k), \boldsymbol{X}(k-1), \ldots, \boldsymbol{X}(k-l)] \in R^{N\times(lm+m)}$. Here, we take the first-order model, i.e., $\boldsymbol{Z} = [\boldsymbol{X}(k), \boldsymbol{X}(k-1)]$, as an example. A new sample at time k then becomes $z = [x_1(k), \ldots, x_m(k), x_1(k-1), \ldots, x_m(k-1)]^T$.

B. INFLUENCE OF VARIABLE SELECTION ON DPCA MONITORING PERFORMANCE

In monitoring a large-scale process, a distributed monitoring scheme can be employed to reduce the redundancy, in which the process decomposition is the key step and will significantly affect the monitoring performance. In the present study, we analyze the influence of variable selection on DPCA monitoring performance by considering cross-correlation and auto-correlation.

Let a numerical example with two variables x_1 and x_2 be

$$\begin{aligned} x_1(k) &= \phi x_1(k-1) + \varepsilon_1(k) \\ x_2(k) &= \theta x_1(k) + \varepsilon_2(k), \end{aligned} \tag{4}$$

where integer k denotes the time index and $\varepsilon_1(k)$ and $\varepsilon_2(k)$ are the serial and joint independent zero-mean Gaussian random variables with variances $1-\phi^2$ and σ^2, respectively. The auto-correlation and cross-correlation are determined by parameters ϕ and θ, respectively. The fault model can be constructed as follows:

$$\begin{aligned} x_1(k) &= x_{1,N}(k) + \beta_1 f \\ x_1(k-1) &= x_{1,N}(k-1) + \beta_2 f \\ x_2(k) &= x_{2,N}(k) + \beta_3 f, \end{aligned} \tag{5}$$

where $x_{1,N}(k), x_{1,N}(k-1)$ and $x_{2,N}(k)$ represent the standardized data under normal status data and f denotes the fault specified by parameters β_1, β_2, and β_3. The following cases are considered in this study: (i) without auto-correlation or cross-correlation; and (ii) with auto-correlation or cross-correlation.

1) WITHOUT AUTO-CORRELATION OR CROSS-CORRELATION

If $\phi = 0$ and $\theta = 0$, then $x_1(k)$, $x_1(k-1)$ and $x_2(k)$ are uncorrelated. Let $z = [z_1, z_2, z_3]^T = [x_1(k), x_1(k-1), x_2(k)]^T$ denote a measurement, and the covariance matrix of a full DPCA with all the three variables involved can be obtained as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{6}$$

The full DPCA can be obtained as

$$t_1 = z_1, \quad t_2 = z_2, \; t_3 = z_3, \tag{7}$$

which indicates that the three PCs (t_1, t_2, and t_3) are exactly the same as the three variables (z_1, z_2, and z_3). The T^2 statistic of a DPCA model can be constructed as

$$T^2 = \sum_{j=1}^{3} t_j^2 = z_1^2 + z_2^2 + z_3^2 \le \chi_\alpha^2(3). \tag{8}$$

A fault F can lead to a ramp change in the measured variables. Thus, the time required for the T^2 statistic to reach the threshold, i.e., the detection delay (defined in [24]), is $J_f = \sqrt{\frac{\chi_\alpha^2(3)}{\beta_1^2+\beta_2^2+\beta_3^2}}$. The Q statistic is not used because all PCs are retained in the dominant subspace. If a reduced PCA model built on only two variables (the three variables are uncorrelated, without loss of generality, using z_1 and z_2), then the T^2 statistic with two variables (PCs) can be expressed as

$$T_{r2}^2 = z_1^2 + z_2^2 \le \chi_\alpha^2(2), \tag{9}$$

and the detection delay is $J_2 = \sqrt{\frac{\chi_\alpha^2(2)}{\beta_1^2+\beta_2^2}}$. Next, the PCA model with only one variable (without loss of generality, using z_1) is considered. The T^2 statistic can be expressed as follows:

$$T_{r1}^2 = z_1^2 \le \chi_\alpha^2(1). \tag{10}$$

Then, the detection delay is $J_1 = \sqrt{\frac{\chi_\alpha^2(1)}{\beta_1^2}}$. The detection delay J reflects the difficulty of a statistic indicating a fault. A larger delay indicates that the fault is more difficult to detect, whereas a smaller delay indicates better monitoring performance with lower non-detection rate. In this section, the detection delays of using the full DPCA model, the reduced two-variable DPCA model, and the reduced one-variable DPCA model are discussed.

First, if we assume that $\beta_3 = 0$, which means that the fault does not cause change on the third variable, then $J_2 < J_f$ because $\chi_\alpha^2(2) < \chi_\alpha^2(3)$. Under this condition, the reduced two-variable DPCA monitoring model performs better than that of a full DPCA model with all the three variables. Second, the monitoring performance of two-variable DPCA model and one-variable DPCA model is geometrically explained, as presented in Fig. 1(a). In Fig. 1(a), the dashed blue circle denotes the control limit for the DPCA monitoring with $x_1(k)$ and $x_2(k)$ (or $x_1(k-1)$), and the dashed red lines denote the control limits used for monitoring each variable individually. Various faults can be generated by the fault parameters β_1, β_2, and β_3. When a fault occurs along the $F1$ direction (yellow area), the DPCA model with only one variable $x_1(k)$ will perform better because $\chi_\alpha^2(1) < \chi_\alpha^2(2)$. Thus, the inclusion of the variable $x_2(k)$ or the time-lagged variable $x_1(k-1)$ in the DPCA model is unnecessary. When a fault occurs along the $F2$ direction (green area), the DPCA model with only $x_2(k)$ or $x_1(k-1)$ will perform better. Thus, the inclusion of $x_1(k)$ in the model is unnecessary. When a fault occurs along the $F3$ direction (gray area), the PCA model with $x_1(k)$ and $x_2(k)$ will provide good monitoring performance.

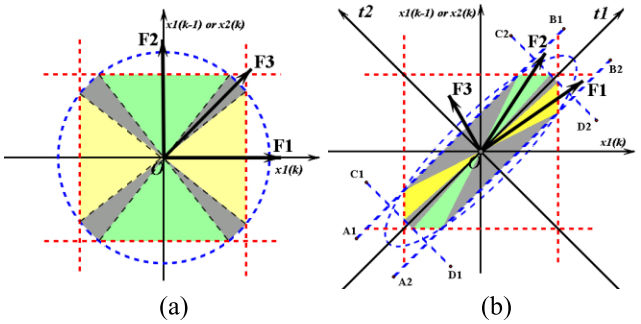

FIGURE 1. DPCA monitoring: (a) uncorrelated case, (b) correlated case.

2) WITH AUTO-CORRELATION OR CROSS-CORRELATION

If $\phi \neq 0$ but $\theta = 0$, then the process has auto-correlation but no cross-correlation. Assume that z_1 and z_2 are correlated with a covariance a_{12} and z_3 is not correlated with the other two variables, i.e.,

$$\begin{aligned} a_{12} &= a_{21} = \mathrm{cov}(z_1, z_2) \\ a_{13} &= a_{31} = \mathrm{cov}(z_1, z_3) = 0 \\ a_{23} &= a_{32} = \mathrm{cov}(z_2, z_3) = 0. \end{aligned} \tag{11}$$

The covariance matrix of the DPCA can be expressed as

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & a_{12} & 0 \\ a_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{12}$$

After PCA transformation, PC scores can be obtained as

$$\begin{aligned} t_1 &= \frac{1}{\sqrt{2}} z_1 + \frac{1}{\sqrt{2}} z_2 \\ t_2 &= z_3 \\ t_3 &= \frac{1}{\sqrt{2}} z_1 - \frac{1}{\sqrt{2}} z_2. \end{aligned} \tag{13}$$

The T^2 statistic of the DPCA model with three variables can be obtained as

$$T_f^2 = \frac{z_1^2 + z_2^2 - 2a_{12}z_1z_2}{1 - a_{12}^2} + z_3^2 \le \chi_\alpha^2(3). \tag{14}$$

The detection delay of the full DPCA model is $J_f = \sqrt{\frac{(1-a_{12}^2)\chi_\alpha^2(3)}{\beta_1^2+\beta_2^2-2a_{12}\beta_1\beta_2}}$. The T^2 statistic of a reduced DPCA using z_1 and z_2 can be given as

$$T_{r2}^2 = \frac{z_1^2 + z_2^2 - 2a_{12}z_1z_2}{(1 - a_{12}^2)} \le \chi_\alpha^2(2), \tag{15}$$

and the detection delay is $J_2 = \sqrt{\frac{(1-a_{12}^2)\chi_\alpha^2(2)}{\beta_1^2+\beta_2^2-2a_{12}\beta_1\beta_2}}$. $\chi_\alpha^2(2) < \chi_\alpha^2(3)$, and in this situation, J_2 will be generally smaller than J_f, and the reduced DPCA model will provide better performance if we assume that $\beta_3 = 0$.

Then, the monitoring performance of two-variable DPCA model and one-variable DPCA model is geometrically explained. The first PC will be used to generate T^2 statistic, whereas the second PC will be left in the residual space.

The Q statistic is constructed to monitor the variations in the residual subspace. The monitoring chart using T^2 and Q in DPCA monitoring is presented in Fig. 1(b), which geometrically illustrates the DPCA monitoring. In Fig. 1(b), the dashed blue ellipse denotes the control limit of T^2 with $x_1(k)$ and $x_2(k)$ (or $x_1(k-1)$), the parallel lines C_1D_1 and C_2D_2 denote the control limit of Q, the parallel lines A_1B_1 and A_2B_2 denote the control limit of T^2 with only one PC retained, and the dashed red lines denote the control limits used for monitoring each variable individually. Various faults can be generated by the fault parameters β_1, β_2, and β_3. When the fault occurs along the $F1$ direction (yellow area), the PCA model with only $x_1(k)$ will perform better because the confidence limit of $x_1(k)$ will be reached earlier. When a fault occurs along the $F2$ direction (green area), the PCA model with only $x_2(k)$ or $x_1(k-1)$ will perform better. When the fault occurs along the $F3$ direction (gray area), the monitoring model with $x_1(k)$ and $x_2(k)$ will provide better monitoring performance. The gray area is the largest among the three areas, indicating that the inclusion of the two variables in one DPCA model is important.

In some situations, $\phi \neq 0$ and $\theta \neq 0$, which means that all variables are correlated with one another. In this situation, no fault irrelevant variable usually exists because the correlation between the variables will provide useful information to detect a fault influencing any of the variables. Based on the above analysis, the existence of irrelevant variables may cause redundancy in the monitoring and may degrade the monitoring performance, which requests that a reduced monitoring model should be established. Determining a fault irrelevant variable through mathematical analysis is difficult because the process model is usually unavailable. In the current work, a stochastic optimization-based method is employed. This method eliminates the fault irrelevant variables based on totally processed data.

C. BAYESIAN FAULT DIAGNOSIS

Bayesian diagnosis identifies the underlying process status based on the currently obtained evidence $\boldsymbol{e}^c$ and historical data $\boldsymbol{D}$. According to Bayes rule, the posterior probability of each possible process status can be calculated as follows [20]:

$$p(F|\boldsymbol{e}^c, \boldsymbol{D}) = \frac{p(\boldsymbol{e}^c|F, \boldsymbol{D})p(F|\boldsymbol{D})}{\sum_F p(\boldsymbol{e}^c|F, \boldsymbol{D})p(F|\boldsymbol{D})}, \tag{16}$$

where $p(\boldsymbol{e}^c|F, \boldsymbol{D})$ is the likelihood of evidence $\boldsymbol{e}^c$ under process status F obtained from historical data and $p(F|\boldsymbol{D})$ is the prior probability of fault status F. The process status with the largest posterior probability can be determined as the underlying process status based on the maximum a posteriori (MAP) principle. To obtain the likelihood from the process history, a marginalization-based solution that involves prior knowledge is provided as follows [22], [23]:

$$p(\boldsymbol{e}_i|F_j, \boldsymbol{D}) = \frac{n(\boldsymbol{e}_i|\boldsymbol{D}_{F_j}) + \alpha\left(\boldsymbol{e}_i|\boldsymbol{D}_{F_j}\right)}{\sum_{h=1}^{K} n(\boldsymbol{e}_h|\boldsymbol{D}_{F_j}) + \sum_{h=1}^{K} \alpha(\boldsymbol{e}_h|\boldsymbol{D}_{F_j})}, \tag{17}$$

where $n(\boldsymbol{e}_i|\boldsymbol{D}_{F_j})$ is the number of evidence $\boldsymbol{e}_i$ under fault status F_j and $\alpha\left(\boldsymbol{e}_i|\boldsymbol{D}_{F_J}\right)$is the number of prior samples assigned to $\boldsymbol{e}_i$ under fault status F_j. The likelihood accounts prior and historical samples and will converge to the relative frequency determined by the historical data as the number of historical samples increases [23]. Meanwhile, due to the existence of the prior sample term, the calculation amount can significantly increase with the number of evidence sources [20], [19]. Then, the main task in using the Bayesian diagnosis system falls on how to generate efficient fault signature evidence.

III. FAULT-RELEVANT PERFORMANCE-DRIVEN DISTRIBUTED MONITORING SCHEME

A. FAULT-RELEVANT PERFORMANCE-DRIVEN PROCESS DECOMPOSITION

The objective of process decomposition is to decompose a process into sub-blocks to reduce the redundancy while the detectability should not be destroyed. When accurate process model and fault model are available, the decomposition can be conducted through mathematical analysis. However, in practice, the analysis may be difficult because (i) the process model or fault model is usually unavailable and (ii) the analysis will be complex when the number of process variables and fault parameters is large [19]. In practical applications, some faults occur constantly or periodically, and these fault data can be collected from process history. Given these reasons, the current study uses a stochastic optimization approach, i.e., the genetic algorithm (GA), to achieve optimal process decomposition.

In using GA, the first step is to establish the fitness function. Assume that process data of G faults, denoted as $F = \{F_1, F_2, \ldots, F_G\}$, can be obtained from process history. Let the measurement in DPCA monitoring be $\boldsymbol{z} = [z_1, \ldots, z_{2m}]^T = [x_1(k), \ldots, x_m(k), x_1(k-1), \ldots, x_m(k-1)]^T$. In performance-driven process decomposition, the sub-blocks are constructed by selecting a subset of variables for each fault. These variables provide the best description of the fault, and the DPCA monitoring model built on these variables achieves the best possible fault detection performance (defined by non-detection rate, NDR). Meanwhile, the false alarm rate (FAR) should be retained at an acceptable level that can be specified according to practical application requirements. For a specific fault F_i, the fitness function is established as follows [19]:

$$\begin{aligned} &\min_{z_j} \ NDR = \frac{N_{F,N}}{N_F} \times 100\% \\ &s.t. \ FAR = \frac{N_{N,F}}{N_N} \times 100\% \leq CL, \end{aligned} \tag{18}$$

where $N_{F,N}$ and N_F denote the number of non-detected fault points and the number of fault points, respectively; $N_{N,F}$ and N_N denote the number of false alarm points and the number of normal samples, respectively; and CL is the specified maximum FAR allowed in practical application.

$z_1(x_1(k))$	...	$z_m(x_m(k))$	$z_{m+1}(x_1(k-1))$	...	$z_{2m}(x_m(k-1))$
1	...	1	0	...	1

FIGURE 2. Chromosome design for fault-relevant performance-driven process decomposition.

π_1	π_2	...	π_{r_1}	π_{r_1+1}	π_{r_1+2}	...	π_{s-1}	π_s
1	0	...	0	1	0	...	0	1

FIGURE 3. Chromosome design for fault-relevant performance-driven evidence source selection.

After constructing the fitness function, the chromosome in GA is designed, as illustrated in Fig. 2. In the chromosome, each gene (element) is designed to encode a variable. For instance, a "1" indicates that the corresponding variable should be selected, whereas a "0" indicates that the corresponding variable should be removed. Using the temporarily selected variables, a DPCA model can be established. Based on the validation data, the fitness function value can be calculated. The GA continues until the best possible performance is achieved or a stop rule is reached.

B. FAULT-RELEVANT PERFORMANCE-DRIVEN OPTIMAL DESIGN OF BAYESIAN DIAGNOSIS SYSTEM

In using Bayesian diagnosis, the first step is to generate fault signature evidence. For the b-th sub-block, assuming that m_b variables are present in $\boldsymbol{z}_b = \left[z_{b,1}, z_{b,2} \ldots, z_{b,m_b}\right]^T$, the DPCA model can be established in the sub-block. The fault signature evidence can be generated by examining the status of each principal component and the DPCA residual. The scaled i-th component score can be obtained as follows:

$$t_{b,i} = \frac{\boldsymbol{z}_b^T \boldsymbol{p}_{b,i}}{\sqrt{\lambda_{b,i}}}, \tag{19}$$

and the corresponding T^2 statistic of this component $T_{b,i}^2$ can be calculated as follows:

$$T_{b,i}^2 = t_{b,i}^2 \sim \chi^2(1). \tag{20}$$

The evidence $\boldsymbol{e}_b = [\pi_{b,1}, \pi_{b,2}, \ldots, \pi_{b,r_b}, \pi_{b,r_b+1}]$ can be generated by discretizing the statistics as follows:

$$\pi_{b,i} = \begin{cases} 0, & if\ T_{b,i}^2 \le \chi_\beta^2(1) \\ 1, & if\ T_{b,i}^2 > \chi_\beta^2(1) \end{cases} (i = 1, 2 \ldots, r_b), \tag{21}$$

$$\pi_{b,r_b+1} = \begin{cases} 0, & if\ Q_b \le Q_{CL,b} \\ 1, & if\ Q_b > Q_{CL,b}, \end{cases} \tag{22}$$

where r_b denotes the number of retained PCs in the b-th sub-block, Q_b is the Q statistic in the b-th sub-block, and $Q_{CL,b}$ is the control limit of Q_b. Then, for all B sub-blocks, $s = \sum_{b=1}^{B} r_b + B$ bits of evidence sources will be generated, and the total number of possible evidence values will be $K = 2^s$.

The calculation amount can significantly increase with the number of evidence sources, which can eventually exceed the computer capacity. By contrast, the evidence sources have distinct importance in discriminating a fault and the existence of irrelevant evidence sources may degrade the monitoring performance [19]. Determining an appropriate number of the most efficient evidence sources to serve as the input of the Bayesian diagnosis system is thus important. Here, the GA is employed. For fault diagnosis problem, the main index is misclassification rate (MCR), and then the fitness function can be constructed as follows [20]:

$$\min_{\pi_i} MCR = \frac{N_{MC}}{N} \times 100\% \quad s.t.\ s \le s_{\max}, \tag{23}$$

where N_{MC} and N denote the number of wrongly classified points and the number of all considered points, respectively. $s_{\max}$ is the acceptable maximum number of retained evidence sources, which is specified according to computation capacity. The chromosome is designed to include all evidence sources, and the existences of the sources are encoded by the values of the elements in the chromosome, as illustrated in Fig. 3.

The procedures of the proposed monitoring scheme consist of offline modeling and online monitoring as follows:

1) OFFLINE MODELING

1) *Collect historical training data under different process statuses;*
2) *Construct new measurements using the 'time lag shift' method;*
3) *Perform GA-based process decomposition;*
4) *Establish distributed DPCA monitoring models and generate fault signature evidence;*
5) *Establish Bayesian fault diagnosis system.*

2) ONLINE MONITORING

1) *Construct new measurement using the 'time lag shift' method;*
2) *Divide the new measurement as the previously obtained sub-blocks;*
3) *Transform the measurement into fault signature evidence;*
4) *Make decision using the Bayesian diagnosis system.*

IV. CASE STUDIES

A. CASE STUDY ON A NUMERICAL EXAMPLE

A numerical AR(1) process modified from [7] is employed to illustrate the performance of the proposed distributed monitoring scheme, which is as follows:

$$\begin{aligned} \boldsymbol{u}_1(k) &= \boldsymbol{A}_1 \boldsymbol{u}_1(k-1) + \boldsymbol{A}_2 \boldsymbol{w}_1(k-1) \\ \boldsymbol{h}_1(k) &= \boldsymbol{A}_3 \boldsymbol{h}_1(k-1) + \boldsymbol{A}_4 \boldsymbol{u}_1(k-1) \\ \boldsymbol{u}_2(k) &= \boldsymbol{A}_1 \boldsymbol{u}_2(k-1) + \boldsymbol{A}_2 \boldsymbol{w}_2(k-1) \\ \boldsymbol{h}_2(k) &= \boldsymbol{A}_3 \boldsymbol{h}_2(k-1) + \boldsymbol{A}_4 \boldsymbol{u}_2(k-1) \\ \boldsymbol{y}(k) &= \boldsymbol{h}_1(k) + \boldsymbol{h}_2(k) + \boldsymbol{v}(k), \end{aligned} \tag{24}$$

TABLE 1. Monitoring results for faults in the numerical example.

Fault No.	Fault description	Selected variables	NDR by DPCA	NDR by reduced DPCA
F1	A ramp change of $0.02(k-50)$ is introduced to $u_{1,1}$	$z_1, z_2, z_3, z_4, z_9, z_{10}$	0.92	0.24
F2	A ramp change of $0.3(k-50)$ is introduced to $w_{1,1}$	$z_1, z_2, z_3, z_4, z_{10}$	0.3	0.14
F3	A ramp change of $0.05(k-50)$ is introduced to $h_{2,1}$	$z_1, z_2, z_5, z_6, z_{10}, z_{11}$	0.96	0.12

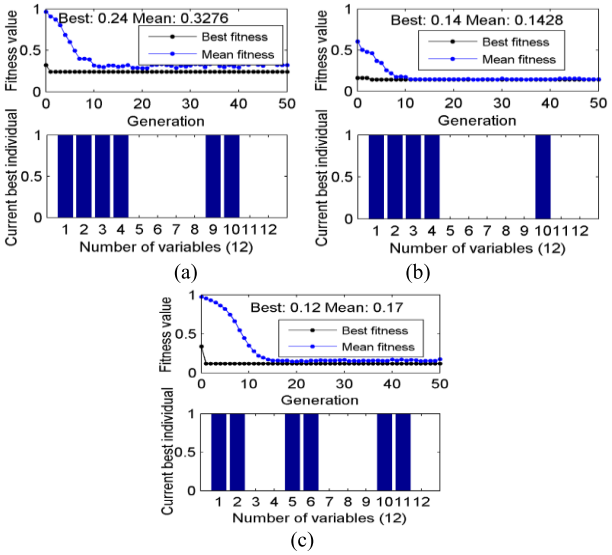

FIGURE 4. Variable selection results for (a) Fault 1, (b) Fault 2, and (c) Fault 3.

where $\boldsymbol{A}_1 = \begin{bmatrix} 0.811 & -0.226 \\ 0.477 & 0.415 \end{bmatrix}$, $\boldsymbol{A}_2 = \begin{bmatrix} 0.193 & 0.689 \\ -0.320 & -0.749 \end{bmatrix}$, $\boldsymbol{A}_3 = \begin{bmatrix} 0.118 & -0.191 \\ 0.847 & 0.264 \end{bmatrix}$, and $\boldsymbol{A}_4 = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$; input $\boldsymbol{w}_i\,(i = 1, 2)$ is the random noise with zero mean and variance 1; and output $\boldsymbol{v}$ is the random noise with zero mean and variance 0.1. The measurement in DPCA model consists of 12 variables as $\boldsymbol{z} = [z_1, \ldots, z_{12}]^T = \left[\boldsymbol{u}_1^T(k)\,\boldsymbol{u}_1^T(k-1)\,\boldsymbol{u}_2^T(k)\,\boldsymbol{u}_2^T(k-1)\,\boldsymbol{y}^T(k)\,\boldsymbol{y}^T(k-1)\right]^T$, where $\boldsymbol{u}_i = \left[u_{i,1}, u_{i,2}\right]^T$ and $\boldsymbol{y}_i = \left[y_{i,1}, y_{i,2}\right]^T$. Three different faults are introduced to the process from the 50 samples listed in Table 1. Together with the normal operating status, a total of four process statuses exist. For each process status, 100 samples are collected to establish the distributed monitoring model. A testing dataset with 200 samples is also generated as F0, F1, F2, and F3, with each status consisting of 50 samples.

The GA variable selection results for the three faults are presented in Fig. 4, the full DPCA monitoring results using all variables are presented in Fig. 5, and the reduced DPCA monitoring results using selected variables are presented in Fig. 6. By comparing Fig. 5 and Fig. 6, the reduced DPCA indicates a fault earlier than that of a full DPCA

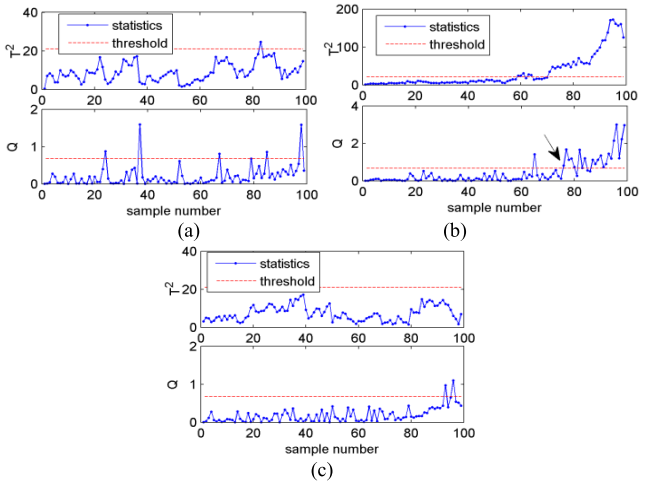

FIGURE 5. Full DPCA fault detection results for (a) Fault 1, (b) Fault 2, and (c) Fault 3.

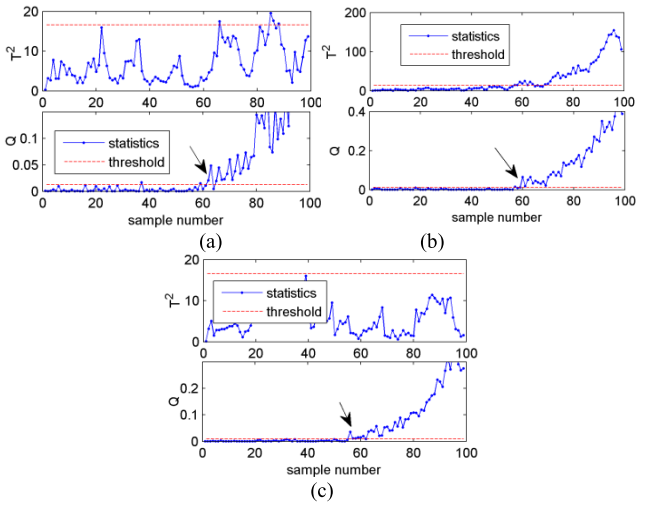

FIGURE 6. Reduced DPCA fault detection results for (a) Fault 1, (b) Fault 2, and (c) Fault 3.

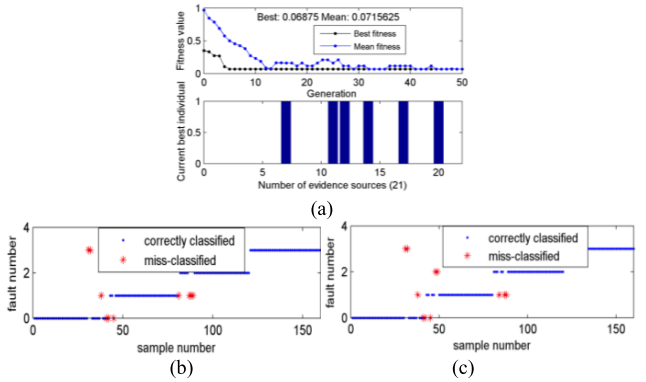

FIGURE 7. (a) Evidence source selection results, (b) diagnosis results using all evidence sources, and (c) diagnosis results using the selected evidence sources.

model (pointed out by the arrows). The performance has been significantly improved as the process decomposition has reduced the redundancy in the monitoring. The NDR for the

TABLE 2. Considered faults in the TE process (NDR).

Fault No.	Fault description	Selected variables	Distributed PCA [15]		Ensemble ICA [26]		Reduced PCA [24]	DPCA	Reduced DPCA
			T^2	Q	I^2	Q	T^2 or Q	T^2 or Q	T^2 or Q
F0	Normal status	-	-	-	-	-	-	-	-
F1 (Fault 1)	A/C feed ratio, B composition constant	$z_7, z_{16}, z_{17}, z_{19}, z_{20}, z_{22}, z_{30}$	0.01	0	0	0	0	0	0
F2 (Fault 5)	Condenser cooling water inlet temp.	$z_1, z_5, z_6, z_9, z_{17}, z_{18}, z_{33}, z_{55}$	0.7	0	0.74	0.69	0	0.52	0
F3 (Fault 10)	C feed temperature (stream 4)	$z_1, z_2, z_5, z_7, z_{18}, z_{19}, z_{20}, z_{31}, z_{51}, z_{52}$	0.43	0.54	0.16	0.17	0.18	0.23	0.12
F4 (Fault 11)	Reactor cooling water inlet temp.	$z_2, z_9, z_{11}, z_{17}, z_{28}, z_{32}, z_{54}, z_{65}$	0.28	0.17	0.6	0.5	0.35	0.06	0.05
F5 (Fault 20)	Unknown	$z_7, z_{11}, z_{13}, z_{21}, z_{64}$	0.48	0.35	0.33	0.33	0.34	0.23	0.14

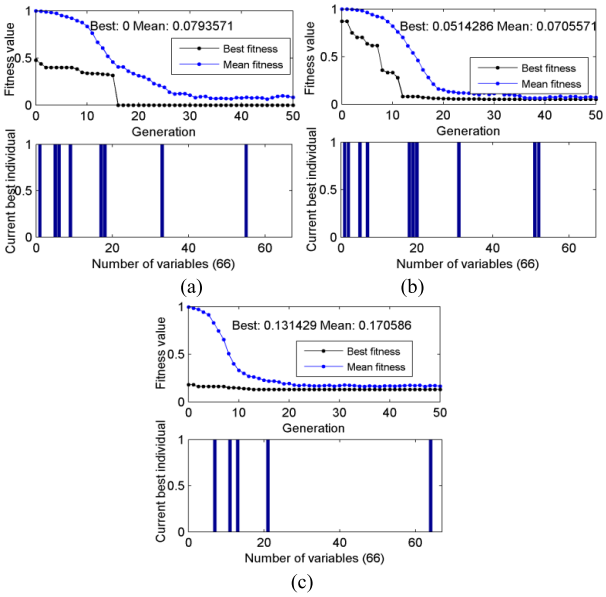

FIGURE 8. GA-based variable selection results for (a) Fault 5, (b) Fault 10, and (c) Fault 20.

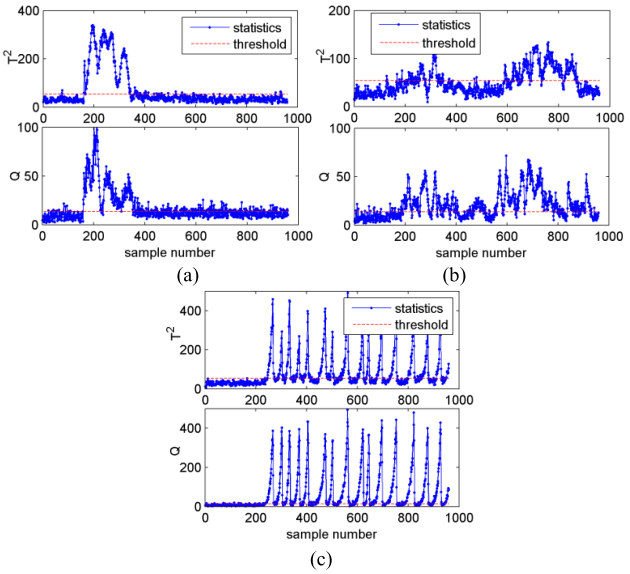

FIGURE 9. Full DPCA fault detection results for (a) Fault 5, (b) Fault 10, and (c) Fault 20.

three faults using the reduced DPCA and the full DPCA are provided in Table I. For establishing the Bayesian diagnosis system, 21 bits of evidence sources are generated. To reduce the calculation amount, GA-based evidence source selection is performed, and the number of evidence sources is reduced to six, as presented in Fig. 7(a). The diagnosis results using all the 21 evidence sources and the selected six evidence sources are presented in Fig. 7(b) and 7(c), respectively. The monitoring performance (MCR) is not destroyed by reducing the sources. However, the calculation amount in Bayesian diagnosis has been reduced from 2^{21} to 2^6, which is important for practical application.

B. CASE STUDY ON THE TE PROCESS

The TE process is a benchmark case designed for testing the process monitoring performance [1], [25]. In the current study, 33 variables and five typical faults are considered [20]. Together with the normal operating condition, a total of six process statuses are used, which are presented in Table 2. The results of Fault 5, Fault 10, and Fault 20 are presented in detail. First, the GA-based variable selection results are presented in Fig. 8, which indicate that the GA optimization converges within 50 generations.

The fault detection results of the full DPCA and the reduced DPCA for the three faults are presented in Fig. 9 and Fig. 10, respectively. Fig. 9(a) shows the full DPCA monitoring results for fault 5, and Fig. 10(a) shows the reduced DPCA monitoring results for fault 5. The NDR has been significantly reduced and the monitoring performance has been significantly improved. The full DPCA and the reduced DPCA monitoring results for fault 10 are presented in Fig. 9(b) and Fig. 10(b), respectively. As shown in the figures, the NDR is also reduced in the reduced DPCA monitoring. Similar results are evident for fault 20, as shown in Fig. 9(c) and Fig. 10(c). The monitoring results (NDR)

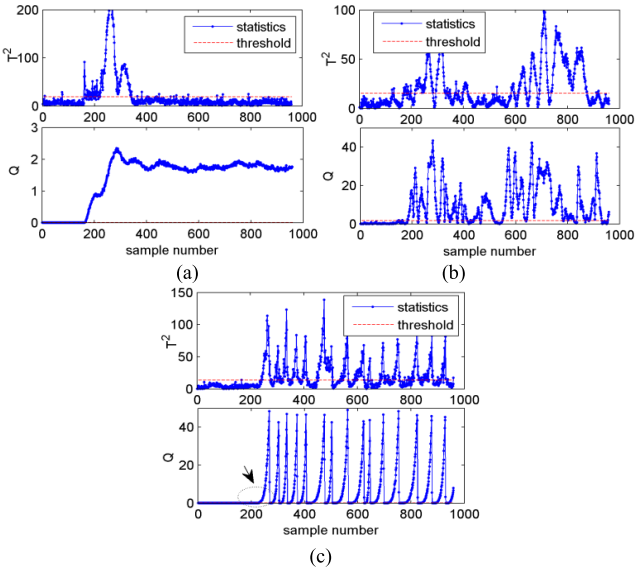

FIGURE 10. Reduced DPCA fault detection results for (a) Fault 5, (b) Fault 10, and (c) Fault 20.

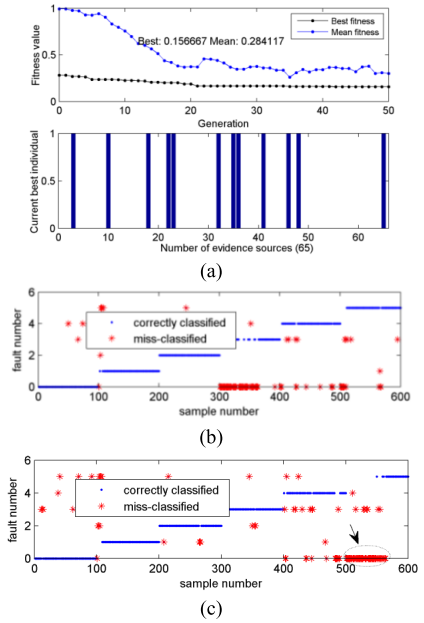

FIGURE 11. (a) GA evidence source selection results, (b) diagnosis results using all evidence sources, and (c) diagnosis results using the selected evidence sources.

of some state-of-the-art methods on the considered faults are presented in the Table 2. The reduced DPCA method generally provides the best monitoring results for the considered five types of faults.

To establish the Bayesian diagnosis system, 65 bits of evidence sources are generated, which will cause significant calculation amount (2^{65}). The GA-based evidence source selection results are presented in Fig. 11(a), in which the maximum number is limited by 12. The fault diagnosis results for the TE training data and testing data are presented in Fig. 11(b) and Fig. 11(c), respectively. Some misclassified points for Fault 20 (marked by the ellipse) are also observed, which are due to the small fault amplitude at the beginning of the fault. The fault can be detected by a DPCA model after approximately 50 points, which is evident in Fig. 10(c) (indicated by the arrow). Most points have been correctly classified, and the diagnosis performance is satisfactory.

V. CONCLUSIONS

In this work, the influence of variable and time-lagged variable selection on the DPCA monitoring performance is analyzed, and a fault-relevant performance-driven distributed DPCA monitoring scheme for large-scale dynamic processes is developed. First, the GA-based fault-relevant variable selection is performed to decompose a large-scale process into several local units and achieve the best possible fault detection performance for each fault. Second, in each unit, a DPCA monitoring model is established to deal with the process dynamics. Then, fault signature evidence is generated, and the Bayesian diagnosis system is established to identify the running-on process status. The proposed monitoring scheme is applied on a numerical example and the TE benchmark process, and the efficiency is demonstrated.

REFERENCES

[1] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. London, U.K.: Springer-Verlag, 2001.
[2] S. X. Ding, "Data-driven design of monitoring and diagnosis systems for dynamic processes: A review of subspace technique based schemes and some recent results," *J. Process Control*, vol. 24, no. 2, pp. 431–449, 2014.
[3] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418–6428, Nov. 2014.
[4] S. J. Qin, "Statistical process monitoring: Basics and beyond," *J. Chemometrics*, vol. 17, nos. 7–8, pp. 480–502, 2003.
[5] Y. Ren and D.-W. Ding, "Fault detection for two-dimensional Roesser systems with sensor faults," *IEEE Access*, vol. 4, pp. 6197–6203, 2016.
[6] J. Chen and K.-C. Liu, "On-line batch process monitoring using dynamic PCA and dynamic PLS models," *Chem. Eng. Sci.*, vol. 57, no. 1, pp. 63–75, 2002.
[7] W. Ku, R. H. Storer, and C. Georgakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 30, no. 1, pp. 179–196, 1995.
[8] G. Li, B. Liu, S. J. Qin, and D. Zhou, "Quality relevant data-driven modeling and monitoring of multivariate dynamic processes: The dynamic T-PLS approach," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2262–2271, Dec. 2011.
[9] A. Negiz and A. Çlinar, "Statistical monitoring of multivariable dynamic processes with state-space models," *AIChE J.*, vol. 43, no. 8, pp. 2002–2020, 1997.
[10] S. W. Choi and I. B. Lee, "Multiblock PLS-based localized process diagnosis," *J. Process Control*, vol. 15, no. 3, pp. 295–306, Apr. 2005.
[11] Q. Liu, S. J. Qin, and T. Chai, "Multiblock concurrent PLS for decentralized monitoring of continuous annealing processes," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6429–6437, Nov. 2014.
[12] S. J. Qin, S. Valle, and M. J. Piovoso, "On unifying multiblock analysis with application to decentralized process monitoring," *J. Chemometrics*, vol. 15, no. 9, pp. 715–742, 2001.
[13] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, "Analysis of multiblock and hierarchical PCA and PLS models," *J. Chemometrics*, vol. 12, no. 5, pp. 301–321, 1998.
[14] Y. Zhang, H. Zhou, S. J. Qin, and T. Chai, "Decentralized fault diagnosis of large-scale processes using multiblock kernel partial least squares," *IEEE Trans. Ind. Informat.*, vol. 6, no. 1, pp. 3–10, Feb. 2010.
[15] Z. Ge and Z. Song, "Distributed PCA model for plant-wide process monitoring," *Ind. Eng. Chem. Res.*, vol. 52, no. 5, pp. 1947–1957, 2013.
[16] Q. Jiang and X. Yan, "Nonlinear plant-wide process monitoring using MI-spectral clustering and Bayesian inference-based multiblock KPCA," *J. Process Control*, vol. 32, pp. 38–50, Aug. 2015.

[17] Q. Jiang and X. Yan, "Just-in-time reorganized PCA integrated with SVDD for chemical process monitoring," *AIChE J.*, vol. 60, no. 3, pp. 949–965, 2014.
[18] Z. Ge and J. Chen, "Plant-wide industrial process monitoring: A distributed modeling framework," *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 310–321, Feb. 2016.
[19] Q. Jiang, X. Yan, and B. Huang, "Performance-driven distributed PCA process monitoring based on fault-relevant variable selection and Bayesian inference," *IEEE Trans. Ind. Electron.*, vol. 63, no. 1, pp. 377–386, Jan. 2016.
[20] Q. Jiang, B. Huang, and X. Yan, "GMM and optimal principal components-based Bayesian method for multimode fault diagnosis," *Comput. Chem. Eng.*, vol. 84, pp. 338–349, Jan. 2016.
[21] B. Huang, "Bayesian methods for control loop monitoring and diagnosis," *J. Process Control*, vol. 18, no. 9, pp. 829–838, 2008.
[22] A. Pernestal, "Probabilistic fault diagnosis with automotive applications," Ph.D dissertation, School of Elect. Eng., KTH Royal Inst. Technol., Stockholm, Sweden, 2007.
[23] F. Qi and B. Huang, "Bayesian methods for control loop diagnosis in the presence of temporal dependent evidences," *Automatica*, vol. 47, no. 7, pp. 1349–1356, 2011.
[24] K. Ghosh, M. Ramteke, and R. Srinivasan, "Optimal variable selection for effective statistical process monitoring," *Comput. Chem. Eng.*, vol. 60, pp. 260–276, Jan. 2014.
[25] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Comput. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, Mar. 1993.
[26] Z. Ge and Z. Song, "Performance-driven ensemble learning ICA model for improved non-Gaussian process monitoring," *Chemometrics Intell. Lab. Syst.*, vol. 123, pp. 1–8, Apr. 2013.

YANG WANG received the B.E. degree from the Department of Automation, Shenyang University of Technology, Shenyang, China, in 2006, and the M.E. degree in measurement and control technology and instrument from Shanghai Maritime University, Shanghai, China, in 2008. She is currently pursuing the Ph.D. degree in control science and engineering with Shanghai University, Shanghai. She is also a Lecturer with Shanghai Dianji University, Shanghai. Her research interests include data-driven fault detection and diagnosis, process monitoring and wireless sensor networks.

QINGCHAO JIANG received the B.E. and Ph.D. degrees from the Department of Automation, East China University of Science and Technology, Shanghai, China, in 2010 and 2015, respectively. In 2015, he was a Post-Doctoral Fellow with the Department of Chemical and Materials Engineering, University of Alberta, Canada. From 2015 to 2016, he was a Humboldt Research Fellow with the Institute for Automatic Control and Complex Systems, University of Duisburg-Essen, Germany. He is currently an Associate Professor with the East China University of Science and Technology, Shanghai. His research interests include data mining and analysis, soft sensor, multivariate statistical process monitoring, and Bayesian fault diagnosis.

JINGQI FU was born in Heilongjiang, China, in 1962. He received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, 1995. He is currently a Professor and a Faculty Member with the School of Mechatronics and Automation, Shanghai University. His research interests include intelligent computing, intelligent sensor, and wireless sensor.