Journal of Process Control 22 (2012) 1567–1581

Contents lists available at SciVerse ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process

Shen Yin[a,b,*], Steven X. Ding[b], Adel Haghani[b], Haiyang Hao[b], Ping Zhang[b]

[a] Institute of Intelligent Control and Systems, Harbin Institute of Technology, P.O. Box 3015, Yikuang Street 2, 150001 Harbin, China
[b] Institute for Automatic Control and Complex Systems, University of Duisburg-Essen, Bismarckstrasse 81 BB, 47057 Duisburg, Germany

ARTICLE INFO

Article history:
Received 22 July 2011
Received in revised form 15 June 2012
Accepted 15 June 2012
Available online 21 July 2012

Keywords:
Process monitoring
Fault diagnosis
Data-driven methods
Tennessee Eastman process

ABSTRACT

This paper provides a comparison study on the basic data-driven methods for process monitoring and fault diagnosis (PM–FD). Based on the review of these methods and their recent developments, the original ideas, implementation conditions, off-line design and on-line computation algorithms as well as computation complexity are discussed in detail. In order to further compare their performance from the application viewpoint, an industrial benchmark of Tennessee Eastman (TE) process is utilized to illustrate the efficiencies of all the discussed methods. The study results are dedicated to provide a reference for achieving successful PM–FD on large scale industrial processes. Some important remarks are finally concluded in this paper.

1. Introduction

PM–FD has been an active research field in the control community during the past several decades. Based on an available process model, a PM–FD system can be successfully designed by a large number of standard methods [2,8,16,40]. Parallel to the research of model-based PM–FD techniques, the so-called data-driven PM–FD methods are currently receiving considerably increasing attention both in application and in research domains. Different from model-based approaches, in which the quantitative model is known a priori, the data-driven PM–FD methods are only dependent on the measured process variables. Thanks to implementation of advanced computer and information technologies, massive amount of measurement data is available, which can be utilized to extract the useful information about current state of the process and support the decision making unit to apply better control and optimization schemes [26]. With their simple forms and less requirements on the design and engineering efforts, the data-driven PM–FD methods become more popular in many industry sectors, especially for large-scale industry applications [6,44]. Recent surveys given by [9,41,42,46,49,50] provide the reader with a comprehensive overview on the basic and advanced data-driven PM–FD schemes. Notice that although data-driven approaches are much simpler than model-based PM–FD techniques, it is still meaningless to directly apply the data-driven approaches on the huge amount of process data, for instance, for a large-scale industrial process including more than thousands of process measurements. In this case, preprocessing is first applied to extract information. Preprocessing includes correlation tests followed by dimension reduction and selection of the variables that explain a significant part of observed process variation. Based on it, efficient data-driven methods can be further selected for PM–FD purpose.

In our recent industrial and research projects dealing with application of data-driven methods for PM–FD, we have noticed that different approaches may show (considerably) different performances even for the same application. Although a number of methods were developed in literature and claimed superior performance on numerous applications, the systematic study and comparison of basic properties have not yet received sufficient attention. These observations motivate us to review the basic data-driven PM–FD methods to understand their original ideas, basic assumptions, implementation conditions as well as limitations. The basic data-driven methods, principle component analysis (PCA), partial least squares (PLS), independent component analysis (ICA), fisher discriminant analysis (FDA), subspace aided approach (SAP) as well as their recent developments, have been considered in this paper. Our aim is to evaluate the applicability and capacity of these methods in the application of industrial processes. For this purpose, all the discussed data-driven PM–FD methods will be applied to an industrial benchmark of Tennessee Eastman (TE) process to illustrate their efficiencies. The contribution of this work is to provide

* Corresponding author at: Institute for Automatic Control and Complex Systems, University of Duisburg-Essen, Bismarckstrasse 81 BB, 47057 Duisburg, Germany.
E-mail addresses: shen.yin@stud.uni-due.de (S. Yin), steven.ding@uni-due.de (S.X. Ding), adel.haghani@uni-due.de (A. Haghani), haiyang.hao@uni-due.de (H. Hao), ping.zhang@uni-due.de (P. Zhang).

0959-1524/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jprocont.2012.06.009

a reference for further PM–FD study on large-scale industrial processes. The concluding remarks on application of the basic PM–FD methods are also summarized in this paper.

The rest of this paper is organized as follows. Section 2 reviews the basic data-driven PM–FD methods as well as their recent developments. A brief comparison on basic assumption, computation complexity and critical design parameters is also presented. Section 3 provides an introduction on the TE process. Based on it, all the discussed methods will be tested and the obtained comparison results are further discussed in Section 4. Finally, the conclusions are presented in the last section.

2. Basic data-driven PM–FD methods

In this section, we would like to review the basic data-driven PM–FD methods as well as their recent developments in the form of off-line design and on-line calculation algorithms. A brief comparison among these methods is also presented.

2.1. Principal component analysis

PCA is a dimensionality reduction technique that preserves the significant variability information in the original data set. Since 1980s, PCA has been successfully applied in numerous areas including data compression, image processing, feature extraction, pattern recognition and process monitoring [24,27]. Due to its simplicity and efficiency in processing huge amount of process data, PCA is recognized as a powerful data-driven PM–FD tool and widely used in practice [10,25,41,42,49].

Consider a process with m measurement signals, which are denoted by a column observation vector, *the off-line design procedure* of standard PCA approach for fault detection purpose can be briefly formulated as:

- *Step 1*: Collect N samples for each measurement and normalize them to zero mean and unit variance, denoted as $Z^T = [\, z_1 \quad \cdots \quad z_N \,] \in \mathcal{R}^{m\times N}$ with the ith normalized observation vector $z_i \in \mathcal{R}^m$, $i = 1, \ldots, N$.
- *Step 2*: Perform singular value decomposition (SVD) on the covariance matrix:

$$\frac{1}{N-1} Z^T Z = P \Lambda P^T, \quad \Lambda = diag(\lambda_1, \ldots, \lambda_m), \quad \lambda_1 \geq \cdots \geq \lambda_m > 0. \tag{1}$$

- *Step 3*: Determine the number of principal components (PCs) l, by a certain criteria in [48], and divide P, Λ into

$$\Lambda = \begin{bmatrix} \Lambda_{pc} & 0 \\ 0 & \Lambda_{res} \end{bmatrix}, \quad \Lambda_{pc} = diag(\lambda_1, \ldots, \lambda_l),$$
$$\Lambda_{res} = diag(\lambda_{l+1}, \ldots, \lambda_m),$$

$$P = [\, P_{pc} \quad P_{res} \,], \quad P_{pc} \in \mathcal{R}^{m\times l}, P_{res} \in \mathcal{R}^{m\times(m-l)}.$$

- *Step 4*: Set thresholds for *SPE* (squared prediction error) [25] and T^2 statistic [47] for a given significant level α:

$$J_{th,SPE} = \theta_1 \left(\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{1/h_0}, \tag{2}$$

$$J_{th,T^2} = \frac{l(N^2-1)}{N(N-l)} F_\alpha(l, N-l) \tag{3}$$

where c_α is the confidence interval that corresponds to the $1-\alpha$ percentile of the normal distribution and can be directly checked from standard tables of the error function,

$$\theta_i = \sum_{j=l+1}^{m} (\lambda_j)^2, \quad i = 1, 2, 3, h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}.$$

An alternative threshold for *SPE* can be found in [38] by utilizing the results from [3].

The on-line computation consists of

- *Step 1*: Normalization of the new measurement sample.
- *Step 2*: On-line computation of *SPE* and T^2 statistic

$$SPE = z^T P_{res} P_{res}^T z, \tag{4}$$

$$T^2 = z^T P_{pc} \Lambda_{pc}^{-1} P_{pc}^T z. \tag{5}$$

- *Step 3*: Fault detection logic according to the following logic

$$SPE \leq J_{th,SPE} \ \ and \ \ T^2 \leq J_{th,T^2} \Rightarrow fault \ \ free, \ \ otherwise \ \ faulty.$$

2.2. Partial least squares

Besides PCA, PLS is another powerful statistical tool and widely used for model building, fault detection and diagnosis purposes [29,30,53]. The basic PLS algorithm, which is implemented with the so-called nonlinear iterative partial least squares algorithm (NIPALS), can be found in [7,19,20]. Suppose that the process under consideration has measurement vector $x \in \mathcal{R}^m$ and a product quality vector under monitoring $y \in \mathcal{R}^a$, *the off-line design procedure* of applying standard PLS approach for PM–FD is formulated as follows:

- *Step 1*: Collect N samples of x and y and normalize them to zero mean and unit variance, denoted as $X^T = [\, x_1 \quad \cdots \quad x_N \,] \in \mathcal{R}^{m\times N}$ and $Y^T = [\, y_1 \quad \cdots \quad y_N \,] \in \mathcal{R}^{a\times N}$.
- *Step 2*: Perform following iterative computations γ times ($k = 1, \ldots, \gamma$):

$$(w_k^*, q_k^*) = arg \max_{\|w_k\|=1, \|q_k\|=1} w_k^T X_k^T Y q_k, \quad X_1 = X,$$

$$t_k = X_k w_k^*, p_k = \frac{X_k^T t_k}{\|t_k\|^2}, \quad X_{k+1} = X_k - t_k p_k^T,$$

$$r_1 = w_1^*, r_k = \prod_{j=1}^{k-1} (I_{m\times m} - w_j^* p_j^T) w_k^*, \quad k > 1$$

where γ is the so-called number of latent variables (LVs) and determined by some certain criteria, e.g. cross validation [54].
- *Step 3*: Store p_k, t_k, q_k, r_k into P, T, Q, R. The correlation model given by standard PLS algorithm is

$$X = TP^T + E,$$
$$Y = TQ^T + F = XM + F, \quad M = RQ^T. \tag{6}$$

- *Step 4*: Set thresholds for the *SPE* and T^2 statistic under a given significant level α

$$J_{th,SPE} = g\chi_\alpha^2(h), \tag{7}$$

$$J_{th,T^2} = \frac{\gamma(N^2-1)}{N(N-\gamma)} F_\alpha(\gamma, N-\gamma) \tag{8}$$

where $g = S/2\mu$ and $h = 2\mu^2/S$, μ and S are respectively the sample mean and variance of *SPE* statistic [38].

The on-line computation for PM–FD consists of

- *Step 1*: Normalization of the new measurement sample.
- *Step 2*: On-line computation of *SPE* and T^2 statistic

$$SPE = \|(I_{m\times m} - PR^T)x\|^2, \tag{9}$$

$$T^2 = x^T R \left(\frac{T^T T}{N-1}\right)^{-1} R^T x. \tag{10}$$

- *Step 3*: Fault detection according to the following logic

$$T^2 > J_{th,T^2} \Rightarrow \textit{faulty in } x, \ \textit{which is related to } y$$

$$SPE > J_{th,SPE} \Rightarrow \textit{faulty in } x, \ \textit{which is unrelated to } y$$

$$T^2 \le J_{th,T^2} \ \textit{and} \ SPE \le J_{th,SPE} \Rightarrow \textit{fault free in } x.$$

2.3. Independent component analysis

ICA is a multivariate statistical tool for extracting the hidden statistically independent components (ICs) from the observed data and originally proposed to solve the signal processing as well as blind source separation problems [17,22,34]. Recently, ICA was applied in the research filed of PM–FD, especially for the process measurement with non-Gaussian distribution [28,32,33,57]. Consider a process with m measurement signals, which are denoted by a column observation vector $x \in \mathcal{R}^m$, *the off-line design procedure* of standard ICA-based fault detection can be briefly summarized as follows:

- *Step 1*: Collect N samples for each measurement and center them to zero mean, which can be written as $X^T = [\, x_1 \ \cdots \ x_N \,] \in \mathcal{R}^{m\times N}$. Calculate $Z^T = QX^T = [\, z_1 \ \cdots \ z_N \,] \in \mathcal{R}^{m\times N}$ with $Q = \Lambda^{-1/2}P^T \in \mathcal{R}^{m\times m}$, where Λ is a diagonal matrix with eigenvalues of covariance matrix $\mathcal{E}(xx^T) \approx 1/(N-1)X^TX$, $P \in \mathcal{R}^{m\times m}$ denotes the related eigenvectors.
- *Step 2*: Perform following iterative computations m times ($k = 1, \ldots, m$):

$$b_k = \underset{\forall b_k, \mathcal{E}(yy^T)=I}{\arg\max} \ (J(y)), \tag{11}$$

$$J(y) \approx [\mathcal{E}\{G(y)\} - \mathcal{E}\{G(v)\}]^2, \quad y = b_k^T z \tag{12}$$

where $J(y)$ is the so-called non-Gaussian measurement function, v is a Gaussian variable with zero mean and unit variance, G is a non-quadratic function [21,23]. Store all b_k into $B = [\, b_1 \ \cdots \ b_m \,] \in \mathcal{R}^{m\times m}$ and calculate demixing matrix $W = B^TQ$.
- *Step 3*: Determine the number of ICs, d, by a certain criterion as listed in [33] with the associated demixing matrix $W_d \in \mathcal{R}^{d\times m}$ and the residual parts in W denoted as $W_e \in \mathcal{R}^{(m-d)\times m}$. Construct following test statistics:

$$I^2 = x^T W_d^T W_d x, \tag{13}$$

$$I_e^2 = x^T W_e^T W_e x, \tag{14}$$

$$SPE = e^T e. \tag{15}$$

- *Step 4*: Set thresholds J_{I^2}, $J_{I_e^2}$ and J_{SPE} for indices (13)–(15) by kernel density estimation (KDE) [36,45].

The on-line computation for PM–FD consists of

- *Step 1*: Center the mean of the new measurement sample.
- *Step 2*: On-line computation of I^2, I_e^2 and *SPE* indices (13)–(15).
- *Step 3*: Fault detection according to the following

$$SPE \le J_{th,SPE} \ \textit{and} \ I^2 \le J_{th,I^2} \ \textit{and} \ I_e^2 \le J_{th,I_e^2} \Rightarrow$$

$$\textit{fault free}, \ \textit{otherwise faulty}.$$

2.4. Fisher discriminant analysis

FDA is a dimensionality reduction technique and has been well studied in the fields of multivariate statistic and pattern classification [14,37]. Due to its ability to discriminate among classes of data, FDA is recognized as an efficient tool for fault classification [4,5,18]. In addition, by defining an additional class of data, which represents normal operating conditions, FDA can also be applied for fault detection purpose [6]. Consider a process with p different operating situations, which can be denoted by p classes of data sets collected from the process, *the off-line design procedure* of FDA for PM–FD can be briefly formulated as:

- *Step 1*: Collect all the p classes of data and stack them into $Z \in \mathcal{R}^{N\times m}$, where $N = \sum_{j=1}^{p} n_j$ (n_j is the number of observations in the jth class), m is the number of measurement signals. Normalize all the p classes of data and finally we have $Z_j \in \mathcal{R}^{n_j\times m}$, $j = 1, \ldots, p$.
- *Step 2*: Calculate the within-class-scatter matrix S_w and between-class-scatter matrix S_b:

$$S_w = \sum_{j=1}^{p} S_j, \quad S_j = \frac{1}{n_j} Z_j^T Z_j, \tag{16}$$

$$S_b = \sum_{j=1}^{p} (\mu_j - \mu)(\mu_j - \mu)^T \tag{17}$$

where $\mu \in \mathcal{R}^m$ and $\mu_j \in \mathcal{R}^m$ denote the mean vectors of stacked matrix Z and the original jth class of data, respectively.
- *Step 3*: Solve the following generalized eigenvalue problem

$$S_b w_k = \lambda_k S_w w_k. \tag{18}$$

In case of inversable S_w (18) is equivalent to solve

$$S_w^{-1} S_b w_k = \lambda_k w_k. \tag{19}$$

Since $rank(S_b) \le p-1$, there exist maximal $p-1$ eigenvectors related to non-zero eigenvalues. Denote a as the number of non-zero eigenvalues, store the related eigenvectors in $W_a = [\, w_1 \ \cdots \ w_a \,] \in \mathcal{R}^{m\times a}$.
- *Step 4*: Set threshold for T^2 statistic under a given significant level α for the jth class:

$$J_{th,T^2}^j = \frac{\bar{a}(N^2-1)}{N(N-\bar{a})} F_\alpha(\bar{a}, N-\bar{a}) \tag{20}$$

where $\bar{a} \le a$ is the largest integer such that $W_{\bar{a}}^T S_j W_{\bar{a}}$ is a full rank matrix with $W_{\bar{a}} = [\, w_1 \ \cdots \ w_{\bar{a}} \,] \in \mathcal{R}^{m\times \bar{a}}$.

The on-line computation consists of

- *Step 1*: Normalization of the new measurement sample.
- *Step 2*: On-line computation of T^2 statistic for the jth class

$$T_j^2 = z^T W_{\bar{a}} (W_{\bar{a}}^T S_j W_{\bar{a}})^{-1} W_{\bar{a}}^T z. \tag{21}$$

- *Step 3*: Fault classification (detection) according to the following logic

$$T_j^2 < J_{th,T^2}^j \Rightarrow data\ (fault)\ belongs\ to\ the\ j-th\ class.$$

The threshold calculation (20) is based on the assumption that the measurement signals follow multivariate Gaussian distribution.

2.5. Subspace aided approach

Based on the well established model-based fault detection and isolation (FDI) techniques, a large number of standard methods can be utilized for PM–FD purpose if a process model is available. The subspace identification methods (SIMs) are powerful tools for identifying the state space process model directly from process data [15,39,51]. From the application viewpoint, the procedure from the rough process data to the final implementation of a model based PM–FD system consists of three steps: (a) (complete) system identification, (b) PM–FD system design, and (c) on-line implementation of the PM–FD system. Recently, Ding et al. [11] proposed a subspace aided approach (SAP), which offers an efficient way for data-driven design of observer-based PM–FD system without identification of the complete process model. More importantly, this approach can deal with PM–FD issue in dynamic systems [12], in which the applications of the aforementioned data-driven methods are considerably limited due to a wide operating range of measurement signals. Suppose that the process under consideration has input vector $u \in \mathcal{R}^m$ and output vector $y \in \mathcal{R}^a$, *the off-line design procedure* of SAP for PM–FD can be briefly formulated as:

- *Step 1*: Arrange the input and output training data into block Hankel matrices

$$Z = \begin{bmatrix} Y \\ U \end{bmatrix}, Y = \begin{bmatrix} Y(k) \\ \vdots \\ Y(k+s) \end{bmatrix}, U = \begin{bmatrix} U(k) \\ \vdots \\ U(k+s) \end{bmatrix},$$

$$Y(j) = [\, y(j) \quad \cdots \quad y(j+N)\,], U(j) = [\, u(j) \quad \cdots \quad u(j+N)\,]$$

where s and N are integers such that $N \gg s \geq n$.
- *Step 2*: Do SVD on $(1/N)ZZ^T$

$$\frac{1}{N}ZZ^T = U_z \begin{bmatrix} \Lambda_{XU} & 0 \\ 0 & \Lambda_\Phi \end{bmatrix} U_z^T, \quad U_z = [\, U_{z,XU} \quad U_{z,res}\,]$$

where Λ_{XU} includes all the singular values, which correspond to the influence of the data set U on the process variables, hence are significant larger than the singular values in Λ_Φ. $U_{z,res}^T$ spans the parity space [52].
- *Step 3*: Select m vectors from $U_{z,res}^T$ and denote them as $[\, \alpha_{s_i} \quad \beta_{s_i}\,] \in U_{z,res}^T$, $i = 1, \ldots, m$. The ith diagnostic observer (DO) can be constructed as follows:

$$\begin{aligned} z^i(k+1) &= A_{z_i} z^i(k) + B_{z_i} u(k) + L_{z_i} y(k) \\ r_i(k) &= g_{z_i} y(k) - c_{z_i} z^i(k) - d_{z_i} u(k) \end{aligned} \tag{22}$$

where

$$A_{z_i} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathcal{R}^{s\times s}, L_{z_i} = -\begin{bmatrix} \alpha_{s_i,0} \\ \alpha_{s_i,1} \\ \vdots \\ \alpha_{s_i,s-1} \end{bmatrix},$$

$$\alpha_{s_i} = [\, \alpha_{s_i,0} \quad \cdots \quad \alpha_{s_i,s}\,], \beta_{s_i} = [\, \beta_{s_i,0} \quad \cdots \quad \beta_{s_i,s}\,] \in \mathcal{R}^{(s+1)m},$$

$$B_{z_i} = [\, \beta_{s_i,0}^T \quad \beta_{s_i,1}^T \quad \cdots \quad \beta_{s_i,s-1}^T\,], d_{z_i} = \beta_{s_i,s}^T,$$

$$g_{z_i} = \alpha_{s_i,s}, c_{z_i} = [\, 0 \quad \cdots \quad 0 \quad 1\,] \in \mathcal{R}^s$$

- In case of Gaussian distributed noise, set the threshold as

$$J_{th} = \chi_\alpha^2(m). \tag{23}$$

Otherwise, KDE technique can be utilized for threshold calculation.

The on-line computation consists of

- *Step 1*: On-line computation of T^2 statistic

$$T^2 = r^T(k)\Lambda_r^{-1} r(k) \tag{24}$$

where $r^T(k) = [\, r_1(k) \quad \cdots \quad r_m(k)\,]$, Λ_r denotes the variance of $r(k)$ that can be determined by Λ_Φ.
- *Step 2*: Fault detection according to the following

$$T^2 < J_{th} \Rightarrow fault\ free, \quad otherwise\ faulty.$$

2.6. A comparison on basic data-driven methods

The basic assumption for applying standard PCA for PM–FD is that the measurement signals follow multivariate Gaussian distribution. Based on generalized likelihood ratio (GLR) test on process measurement [1], a modified statistic was proposed in [10] which delivers an optimal fault detection under given confidence level. The issues related to fault isolation and identification have also been discussed therein. To deal with autocorrelation of process variable, the so-called dynamic PCA (DPCA) has been proposed in [31]. Suppose that the observations on the time interval $[k-N, k]$ are available, the data matrix can be formed in the following manner,

$$Z_k(h) = \begin{bmatrix} z_k^T(h) \\ \vdots \\ z_{k+h-N}^T(h) \end{bmatrix}, z_i^T(h) = [\, z_i^T \quad \cdots \quad z_{i-h}^T\,]$$

where $i = k-N, \ldots, k$. The remaining procedures for off-line design and on-line computation are identical with standard PCA. Notice that the PCA and DPCA are standard approaches in the framework of multivariate statistical process monitoring scheme, in which the normalization procedure plays a central role to construct test statistics for fault detection purpose. Although the data matrix of DPCA also contains time delayed vectors, the normalization procedure cannot deal with the wide operating regions of process variables. From the fault detection point of view, PCA and DPCA are not different as generalized likelihood ratio test with normalized process

measurement. On the other hand, the basic idea of SAP is to construct a parity space based residual generator directly from process data. From the fault detection point of view, instead of normalization procedure in PCA and DPCA, subspace aided approach uses singular value decomposition to remove the deterministic influence in order to construct residual signal or test statistic for process monitoring.

The original idea behind the PLS is to identify the correlation model (6) by utilizing covariance information $cov(x, y)$ and based on it, to predict y using the (online) observation x. On the assumption that x and y follow multivariate Gaussian distribution, fault detection can be achieved through suitable test statistics based on $\hat{x} = PR^T x$ and the residual $\tilde{x} = (I_{m\times m} - PR^T)x$, see (7)–(10). Since PLS related approach is aiming to detect the faults in process variables that are mostly related to product quality variable, the quality variable will not be directly used in on-line implementation as shown in (9) and (10). Although the PLS-based PM–FD technique works in many applications, it has been proven in [35] that standard PLS performs an oblique decomposition on measurement space thus $\hat{x}$ may contain variations orthogonal to y that are not useful for prediction, while residual $\tilde{x}$ may have large variations that hamper overall efficacy of the process monitoring scheme. To solve this problem, Zhou et al. [58] proposed the so-called total projection to latent structure (TPLS) approach, which is based on the results of standard PLS algorithm and makes further decomposition on certain subspaces. An alternative modified approach (MPLS) was recently proposed in [56], which firstly estimates the correlation model in the least-square sense and based on it, further performs an orthogonal decomposition on measurement space. The modified approach does not only deliver a better PM–FD performance but also requires less computation in comparison with all the existing PLS-based PM–FD approaches.

Compared with all the other methods, the calculation involved in ICA is more complicated. The basic assumption of ICA is the measurement signal can be described as a linear combination of non-Gaussian variables, i.e. ICs. Based on it, the process measurement x follows non-Gaussian distribution and the thresholds for (13)–(15) cannot be determined by F-distribution or χ^2 test. Thus, KDE technique, which provides a non-parametric way of estimating the probability density function, is widely utilized in literature to set appropriate thresholds for PM–FD purpose. The recently proposed modified ICA (MICA) algorithm offers a unique solution of ICs and also reduces the computation load compared with the standard approach [32].

Table 1 offers a brief comparison among all the discussed data-driven methods, in which the basic assumption on data, computation complexity and critical design parameters are mainly taken into consideration. One basic assumption for successful implementation of PCA, PLS and FDA related approaches is that the process data follow multivariate Gaussian distribution. In addition, FDA is comparable with PCA and the data sets should be well documented in order to offer detailed information about normal operating condition and complete faulty cases. Since ICA assumes that the process measurement is a linear combination of ICs (non-Gaussian) and thus does not follow Gaussian distribution. The basic PCA, PLS, ICA and FDA related methods are mainly used for the applications in the steady state, while SAP is suitable to cope with PM–FD issue in dynamic processes. Moreover, SAP does not have any special assumption on the process data. The χ^2 test and *KDE* can be used for threshold computation in case of Gaussian and non-Gaussian distributed noise, respectively.

A brief computation complexity analysis is also listed in Table 1, in which ICA related algorithms are the most complicated to solve iterative constraint optimization problems (11) and (12). Except ICA/MICA, the computation burden of the other methods mainly come from performing SVD on covariance or correlation matrices with different dimensions. SAP and DPLS have higher computation cost than standard PCA, since SVD is implemented on higher dimensional Hankel matrices. The core of FDA is to solve a generalized eigenvalue decomposition (EVD) problem, which is a little more complex compared with standard PCA. In addition, MPLS has a comparable computation cost as PCA and seems simpler than PLS and TPLS. Consider that the SAP does not need the normalization step, PCA, MPLS and SAP have relatively lower computational cost over all the methods.

It is worth mentioning that the numbers of PCs, ICs and LVs are important design parameters in PCA, ICA and PLS related methods to achieve successful PM–FD. The leave-N-out cross validation based PRESS statistic [48,55] is mostly referred in the literature for selecting the numbers of PCs and LVs. Although there is no standard criterion to calculate the number of ICs, some methods are suggested in [32,33]. In [32], the authors suggested to set the number of ICs as the same as the number of PCs for a fair comparison purpose. For SAP, the number of s can be determined according to the criteria utilized in [11]. The further discussion about the influences of design parameters on PM–FD performance will be presented based on the simulation results of TE process.

3. TE benchmark process

In this section, we would like to briefly introduce an industrial benchmark of TE process. Based on the well-established benchmark process, all the discussed methods will be further applied to demonstrate their efficiencies. TE process model is a realistic simulation program of a chemical plant which is widely accepted as a benchmark for control and monitoring studies. The process is described in [13] and the FORTRAN code of the process is available over internet. Fig. 1 shows the flow diagram of the process with five major units, i.e. reactor, condenser, compressor, separator and stripper. The process has two products from four reactants. Additionally, an inert and a by-product are also present making a total of 8 components denoted as A, B, C, D, E, F, G and H. The process allows total 52 measurements out of which 41 are of process variables and 11 are manipulated variables, see Tables 13 and 14 in Appendix A. Downs and Fogel [13] initially defined 20 process faults and an additional valve fault

Table 1
A brief comparison among basic data-driven methods.

Method	Assumption on data	Computation complexity	Parameter
PCA	Multivariate Gaussian distribution	Low: 1 SVD on $m \times m$ matrix	No. of PCs
DPCA	Same as PCA	Medium: 1 SVD on $hm \times hm$ matrix	No. of PCs, h
FDA	Same as PCA, well documented data sets	Medium: generalized EVD on $m \times m$ matrix	no
PLS	Same as PCA, clear input–output relationship	Medium: γ times SVD on $m \times m$ matrix	No. of LVs
TPLS	Same as PLS	Medium: cost of PLS + 2 SVD on $m \times m$ +1 SVD on $\alpha \times \alpha$ matrix	No. of LVs
MPLS	Same as PLS	Low: 2 SVD on $m \times m$ matrix	no
ICA/MICA	Measurement is a linear combination of ICs	High: cost of PCA + iterative constraint optimization problems	No. of ICs
SAP	Clear input–output relationship	Medium: 1 SVD on $s(\alpha + m) \times s(\alpha + m)$ matrix	No. of s

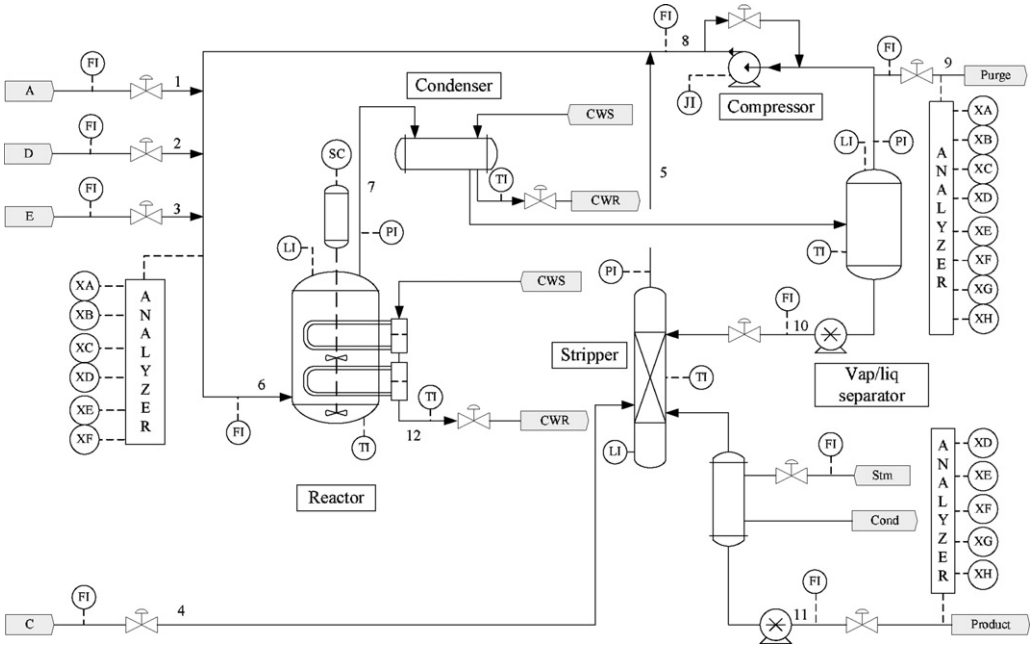


Fig. 1. The Tennessee Eastman process.

was further introduced in [6], see Table 2. As no prior knowledge about the mathematical model of TE process is available, the PM–FD system shall be designed only based on the process data.

Table 2
Descriptions of process faults in TE process.

Fault number	Process variable	Type
IDV(1)	A/C feed ratio, B composition constant	Step
IDV(2)	B composition, A/C ration constant	Step
IDV(3)	D feed temperature	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss	Step
IDV(7)	C header pressure loss-reduced availability	Step
IDV(8)	A, B, and C feed composition	Random variation
IDV(9)	D feed temperature	Random variation
IDV(10)	C feed temperature	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	Unknown	Unknown
IDV(17)	Unknown	Unknown
IDV(18)	Unknown	Unknown
IDV(19)	Unknown	Unknown
IDV(20)	Unknown	Unknown
IDV(21)	The valve fixed at steady state position	Constant position

Table 3
Design parameter selection.

Approaches	PCA	DPCA	ICA	MICA	PLS	TPLS	SAP
Design parameters	PCs = 9	PCs = 17	ICs = 9	ICs = 9	LVs = 6	LVs = 6	$s = 13$

The data sets given in [6] are widely accepted for PM–FD study, in which 22 training sets (including normal operation condition) were collected to record the process measurements for 24 operation hours. Correspondingly, 22 generated (on-line) test data sets were generated including 48 h plant operation time, in which the faults were introduced after 8 simulation hours. By considering the time constants of the process in closed loop, the sampling time was selected as 3 min. These data sets can be downloaded from http://brahms.scs.uiuc.edu.

According to the original TE code, a Simulink code provided by the Ricker [43] is available to simulate the plant's closed-loop behavior. Based on the simulator, the operation modes, measurement noise, sampling time and magnitudes of the faults can be easily modified and thus its generated data sets can be more helpful for PM–FD comparison study. Note that the control structure utilized in [43] is different from the one in [6], which may lead some differences in later simulation study. In our analysis, the base operating mode of TE process is considered to be identical with the case in [6]. The simulator can be downloaded from http://depts.washington.edu/control/LARRY/TE/download.html.

4. Comparison study based on TE

All the discussed data-driven PM–FD methods, including PCA, DPCA, PLS, TPLS, MPLS, FDA, ICA, MICA and SAP, will be applied to TE process for a comparison study. Two generally used indices, i.e. fault detection rate (FDR) and false alarm rate (FAR), are mainly considered here for evaluating PM–FD performance [6,32,58].

$$\mathrm{FDR} = \frac{\text{No. of samples } (J > J_{th} | f \neq 0)}{\text{total samples } (f \neq 0)} \times 100$$

$$\mathrm{FAR} = \frac{\text{No. of samples } (J > J_{th} | f = 0)}{\text{total samples } (f = 0)} \times 100$$

Since the faults in TE as well as other industrial processes may occur in any measurement subspaces, which are generally unknown in practice, a reasonable fault detection logic is based on joint use of the related test statistics, i.e. if one of the test statistics exceeds threshold, a successful fault detection is achieved. The

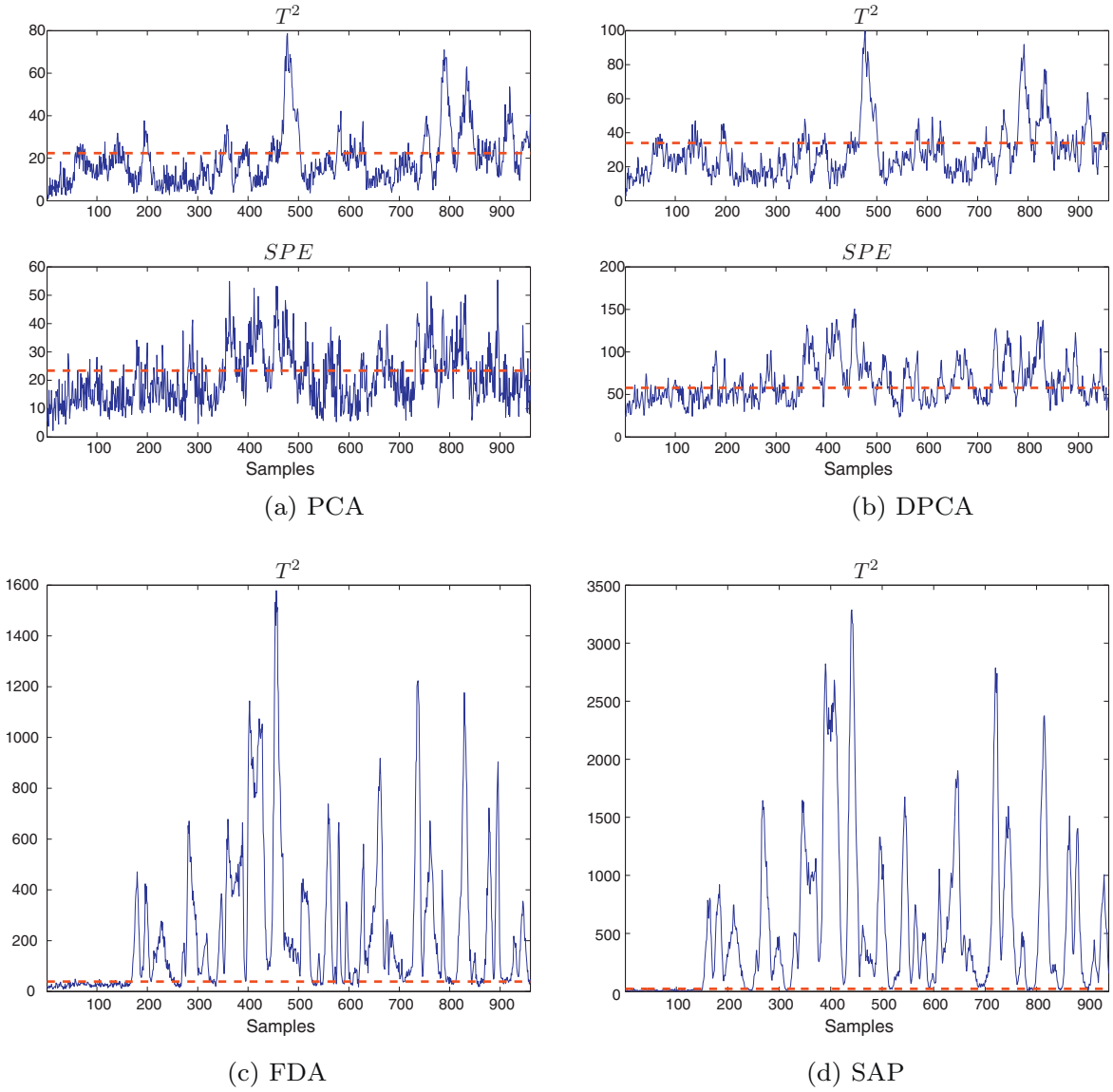


Fig. 2. Process monitoring using PCA, DPCA, FDA, SAP in case of IDV(16).

data sets given in [6] and simulators offered by [43] are utilized in our study in order to achieve convincing results.

4.1. Study on the data sets given in [6]

In this simulation, 22 process measurements (XMEAS(1–22)), 11 manipulated variables (XMV(1–11)) are included for FDA, PCA and ICA related methods. For PLS and SAP, the indicator for component G (XMEAS(35)) is treated as product quality variable (output variable) and other 33 process variables (input variables).

The cross validation based PRESS statistic [48,55] is firstly applied to select the number of PCs and LVs: 9 and 17 PCs are selected for PCA and DPCA, respectively. 9 ICs are selected for ICA and MICA according to [32]. For PLS and TPLS, the number of LVs is selected as 6 based on the cross validation result given by [58]. In addition, s is equal to 13 according to the criteria in [11]. The selected parameters are summarized in Table 3.

Table 4 offers a detailed FDRs by utilizing all the methods on TE process, in which the red color denotes the highest FDR related to a certain type of fault. In the first block of Table 4, i.e. IDV(1–2), IDV(4–8), IDV(12–14) and IDV(17–18), all the methods offer high FDRs except PCA/DPCA and PLS in IDV(5). For the second block, i.e. IDV(10–11), IDV(16) and IDV(19–21), the SAP, MPLS and DPCA, provide superior fault detection performance over all the other methods. For IDV(3), IDV(9) and IDV(15), all methods give low FDRs thus cannot detect the faults successfully. Moreover, Table 5 shows the FARs by applying all the methods in fault free case, from which SAP and ICA related methods give significant better results, while TPLS shows the highest FAR.

Since indicator for component G (XMEAS(35)) is treated as product quality variable, the process faults IDV(3–4), IDV(9,11), IDV(14–15) and IDV(19) have almost no influence on product quality while other faults cause significant variations on quality variable, i.e. with higher severities. The detailed process monitoring figures of two typical faults, i.e. IDV(16) and IDV(19)), have been given in Figs. 2–7 to show the original time trends of each method in order to give insightful features.

In case of IDV(16), the detailed process monitoring figures by using PCA, DPCA, FDA and SAP are shown in Fig. 2, from which it can be clearly seen that the FDA and the SAP are more sensitive than the PCA and DPCA methods. Similarly, the ICA and MICA

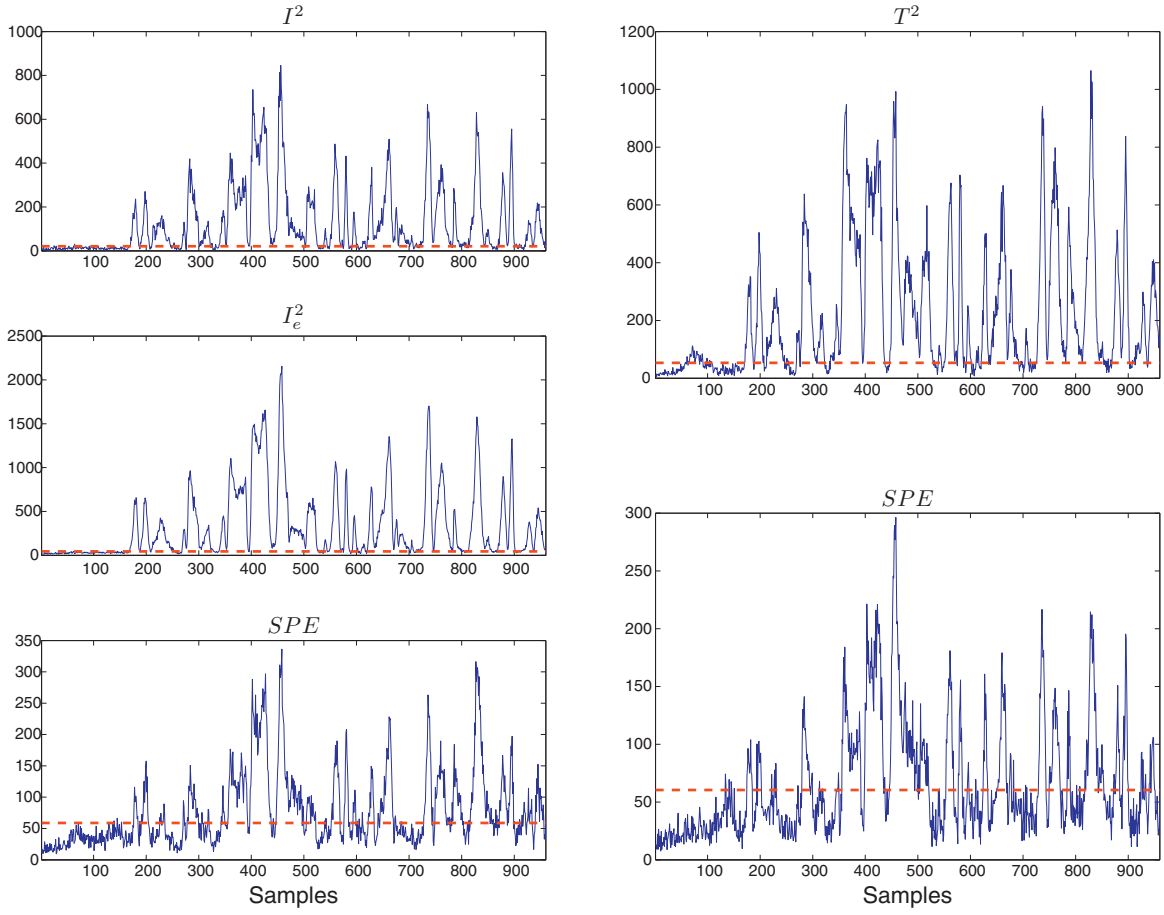


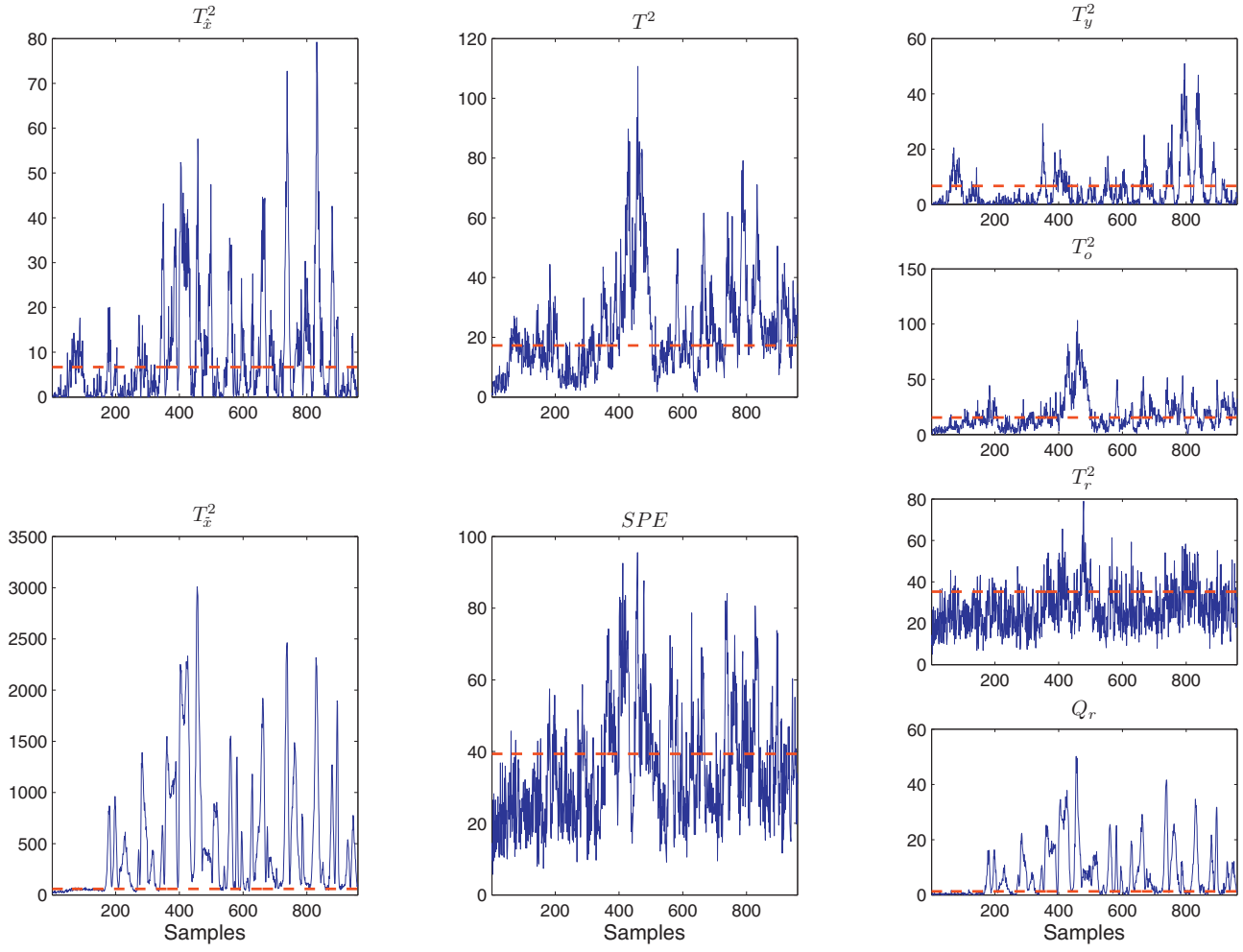
Fig. 3. Process monitoring using ICA, MICA in case of IDV(16).

Table 4
FDRs (%) based on TE data sets given in [6].

Fault	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(1)	99.88	99.88	**100**	99.88	**100**	99.88	99.88	**100**	99.63
IDV(2)	98.75	**99.38**	98.25	98.25	98.75	98.63	98.88	98.88	97.88
IDV(4)	**100**	**100**	**100**	87.63	**100**	99.5	**100**	**100**	99.88
IDV(5)	33.63	43.25	**100**	**100**	**100**	33.63	**100**	**100**	**100**
IDV(6)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(7)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.88
IDV(8)	98	98	98.25	97.63	98.13	97.88	98.5	**98.63**	95.88
IDV(12)	99.13	99.25	**99.88**	**99.88**	99.75	99.25	99.63	**99.88**	**99.88**
IDV(13)	95.38	95.38	95.25	95	95.63	95.25	**96.13**	95.5	94.88
IDV(14)	**100**	**100**	**100**	99.88	**100**	**100**	**100**	**100**	97.63
IDV(17)	95.25	**97.25**	96.88	93	96.63	94.25	96	97.13	97.13
IDV(18)	90.5	90.88	90.5	89.75	90.75	90.75	**91.88**	91.25	91
IDV(10)	60.5	72	89.25	85.88	87.13	82.63	91	91.13	**95.5**
IDV(11)	78.88	**91.5**	78.88	61.63	73.38	78.63	86.13	83.25	84.75
IDV(16)	55.25	67.38	92.38	83.38	83.25	68.38	90.75	94.28	**94.88**
IDV(19)	41.13	87.25	92.88	80.25	87.88	26	82.88	**94.25**	88.5
IDV(20)	63.38	73.75	91.38	86	81.88	62.75	78.38	**91.5**	83.75
IDV(21)	52.13	61	56.38	70.75	52.75	59.88	66.38	**72.75**	38.63
IDV(3)	12.88	12.25	4.5	14.25	7	14.25	**24,25**	18,75	6.38
IDV(9)	8.38	12.88	4.75	8.88	6.25	14.5	**23.5**	12.13	0.88
IDV(15)	14.13	19.75	7.75	10.75	12.63	23	**29.88**	23.25	29.5

Table 5
FARs (%) based on TE data sets given in [6].

Fault Free	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(0)	6.13	10.13	2.75	1.63	6.38	10	19.62	10.75	**1.5**



(a) MPLS (b) PLS (c) TPLS

Fig. 4. Process monitoring using MPLS, PLS, TPLS in case of IDV(16).

methods shown in Fig. 3 give superior results than PCA and DPCA. In addition, the results of PLS related approaches are shown by Fig. 4, from which the MPLS and TPLS provide much better fault detection performance than the standard one.

For fault IDV(19), which has almost no influence on quality variable, the figures of detailed process monitoring are shown in Figs. 5–7, in which the FDA and SAP also offer better results than the PCA and DPCA methods. Moreover, MPLS and TPLS offer not only higher FDRs but also correct fault diagnosis information about the properties of the faults. ICA and MICA are better than standard PCA and PLS, however, have not shown further advantages over the other methods.

As aforementioned, the selection of design parameters, i.e. numbers of PCs, ICs and LVs, may significantly influence the PM–FD performance. Based on this observation, another simulation test is performed, in which the design parameters are selected by different criteria. For standard PCA, 17 PCs are selected according to percent variance test [6] including about 90% variation information and the same number of ICs is selected for ICA/MICA for a fair comparison. The order of time lag in DPCA is determined as $h = 2$ and the number of PCs is 42 which contains 90% of the total variances. For PLS and TPLS, the number of latent variables is selected as 29 based on the leave-one-out cross validation test. In SAP, s is changed to 15. Based on these design parameters as summarized in Table 6, Table 7 offers detailed FDRs of PCA/DPCA, ICA/MICA, PLS/TPLS and SAP approaches. For convenient comparison purpose, the FDRs given by FDA and MPLS, which are identical with the ones in Table 4, are also listed.

According to the FDRs given by Tables 4 and 7, it is obvious that different design parameters will significantly influence PM–FD performance for PCA, DPCA and PLS approaches, especially in cases of IDV(5), IDV(16) and IDV(19–21). However, the design parameter of SAP has little influence on FDRs. In addition, SAP method provides the lowest FAR as shown in Tables 5 and 8. Generally speaking, DPCA and TPLS/MPLS always provide improvements on FDRs compared with standard PCA and PLS approaches, while ICA related approach has not shown evident improvements over other methods.

Table 6
Selection of different design parameters.

Approaches	PCA	DPCA	ICA	MICA	PLS	TPLS	SAP
Design parameters	PCs = 17	PCs = 40	ICs = 17	ICs = 17	LVs = 29	LVs = 29	$s = 16$

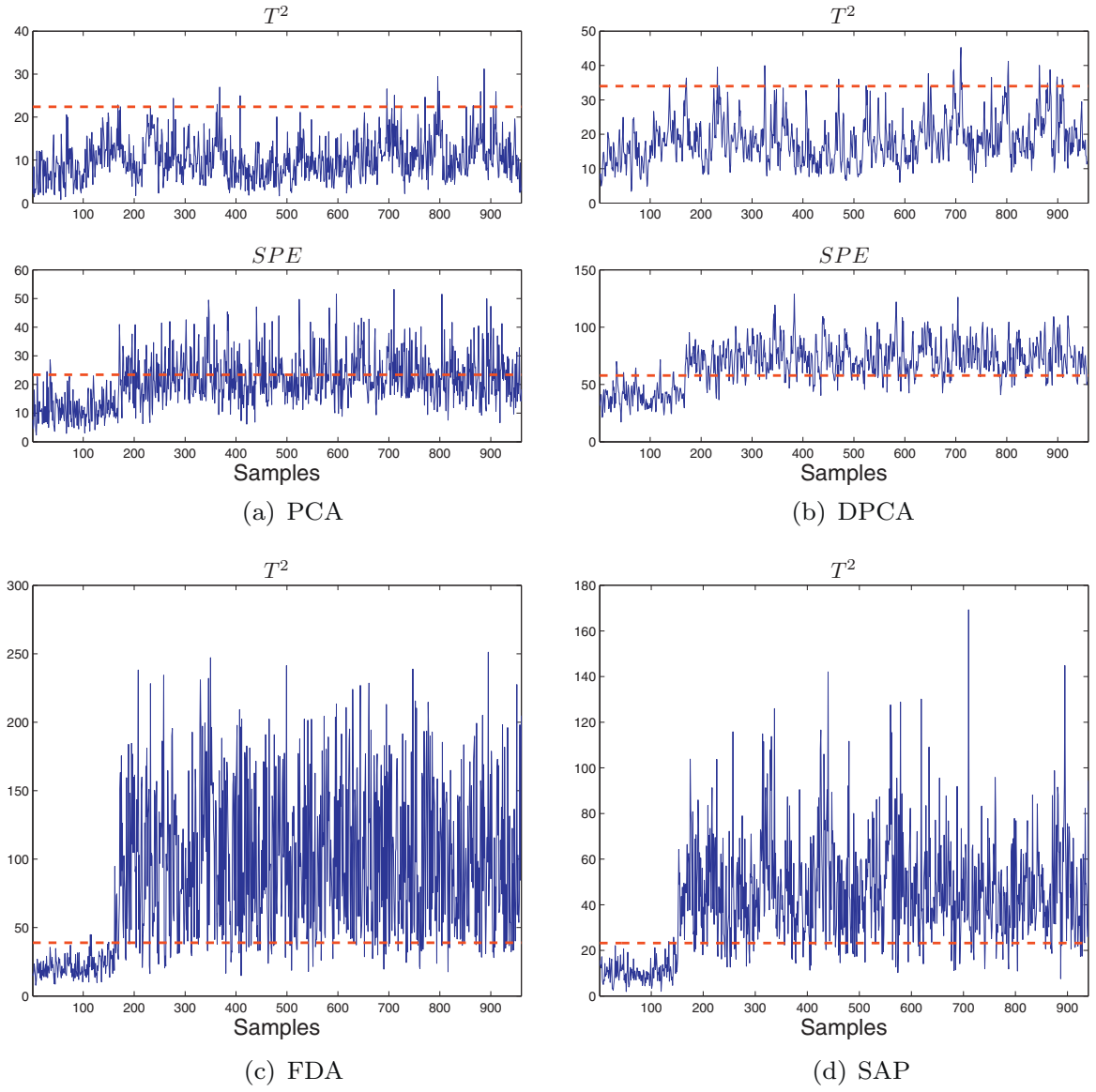


Fig. 5. Process monitoring using PCA, DPCA, FDA, SAP in case of IDV(19).

Table 7
FDRs (%) based on TE data sets given in [6] with different design parameters.

Fault	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(1)	**100**	**100**	99.88	**100**	**100**	**100**	**100**	**100**	99.63
IDV(2)	**99.38**	99.25	98.75	98.38	98.75	98.88	98.88	98.88	98.5
IDV(4)	**100**	**100**	**100**	93.13	**100**	**100**	**100**	**100**	99.63
IDV(5)	34.75	67.75	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(6)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(7)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(8)	**98.63**	98.13	97.88	97.63	98.13	98.5	**98.63**	**98.63**	98.13
IDV(12)	99	99.25	99.88	99.88	99.75	99.88	**100**	99.88	99.88
IDV(13)	95.75	96	95.38	94.88	95.63	95.38	95.63	95.5	**96.13**
IDV(14)	**100**	**100**	**100**	99.88	**100**	**100**	**100**	**100**	97.75
IDV(17)	96.88	**98.13**	96.88	94.5	96.63	97	97.13	97.13	97.25
IDV(18)	91.13	**92.63**	90.5	90	90.75	91	91.25	91.25	91
IDV(10)	71	83.25	89	87.63	87.13	82.63	91.38	92.75	**95.75**
IDV(11)	83	**97.38**	79.75	64.5	73.38	83.38	84.63	83.25	83.88
IDV(16)	65.75	80.13	92.25	88.38	83.25	94.75	95.25	94.38	**97.75**
IDV(19)	47.38	**95.25**	93.13	84	87.88	95	92.5	94.25	88.63
IDV(20)	71.5	80.88	90.88	89	81.88	91.38	91	**91.5**	86.63
IDV(21)	58.13	63.13	55.63	70.75	52.75	64.87	70.5	**72.75**	39.75
IDV(3)	10.25	**23.25**	5	13.25	7	11.75	21,88	18,75	3.13
IDV(9)	9.88	**23.25**	4.88	9.38	6.25	8.38	15.38	12.13	2.38
IDV(15)	17.25	25.88	10.25	12	12.63	22	**28.5**	23.25	15.38

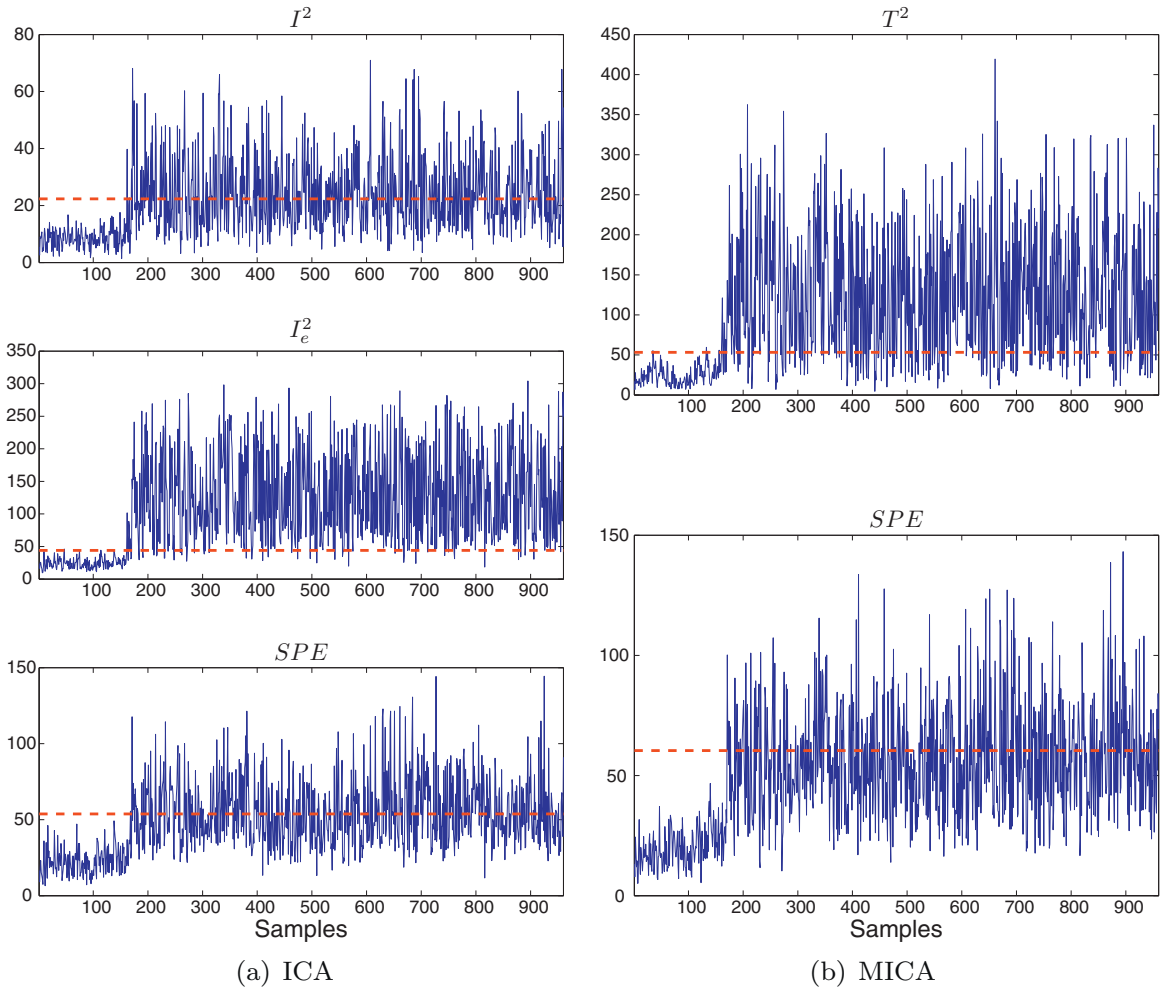


Fig. 6. Process monitoring using ICA, MICA in case of IDV(19).

4.2. Study on the data sets generated from simulator given in [43]

Although different control strategies are implemented in the simulator, the similar operation conditions described in the last subsection are taken for collecting training and (on-line) test data sets, in which 21 training data sets (without IDV(21)) with the corresponding (on-line) test data sets are collected to record the process measurements of 24 and 48 operation hours. To further investigate the ICA approaches, a fault is introduced after 8 h of simulation time, in the form of increased non-Gaussian noise on the process variables. Since the uniform-distributed noise is typical non-Gaussian and widely accepted to investigate the effectiveness of ICA approach [17,28,32–34], the uniform-distributed signal is added on the measurement with the interval related to 0.2 times variance of the associated process variable. The cross validation based PRESS statistic is firstly applied for selecting the numbers of PCs, ICs and LVs, which are listed in Table 3. Since the magnitudes of faults defined in the simulator are very large, the modified magnitudes, which are less than 25% of original values, are implemented in the simulation study. For each type of faults, one hundred Monte Carlo simulations are performed to obtain FDRs of all the discussed methods.

Tables 9 and 10 summarize the detailed FDRs and FARs. In the first block of Table 9, all the tested methods give similar FDRs. The evident difference among FDRs can be found in the second block of this Table, where SAP offers much better FDRs over all the other methods. However, the faults listed in the third block are undetectable by all the given methods. In addition, PCA approach gives the best FARs listed in Table 10.

Another simulation test is performed with different design parameters selected by percent variance test, which are listed in Table 6, and leave-one-out cross validation. The FDRs given by PCA, ICA/MICA and PLS methods are significantly influenced through the parameters change, which can be seen from the second block of Tables 9 and 11. Similar to observations in the second block of Table 9, in Table 11 SAP also gives better FDRs in most cases. In addition, DPCA and TPLS/MPLS are evident to offer better FDRs than standard PCA and PLS approaches. ICA related approaches

Table 8
FARs (%) based on TE data sets given in [6] with different design parameters.

Fault Free	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(0)	6.38	15.13	2.63	1.5	6.38	7.12	12.13	10.75	**1.25**

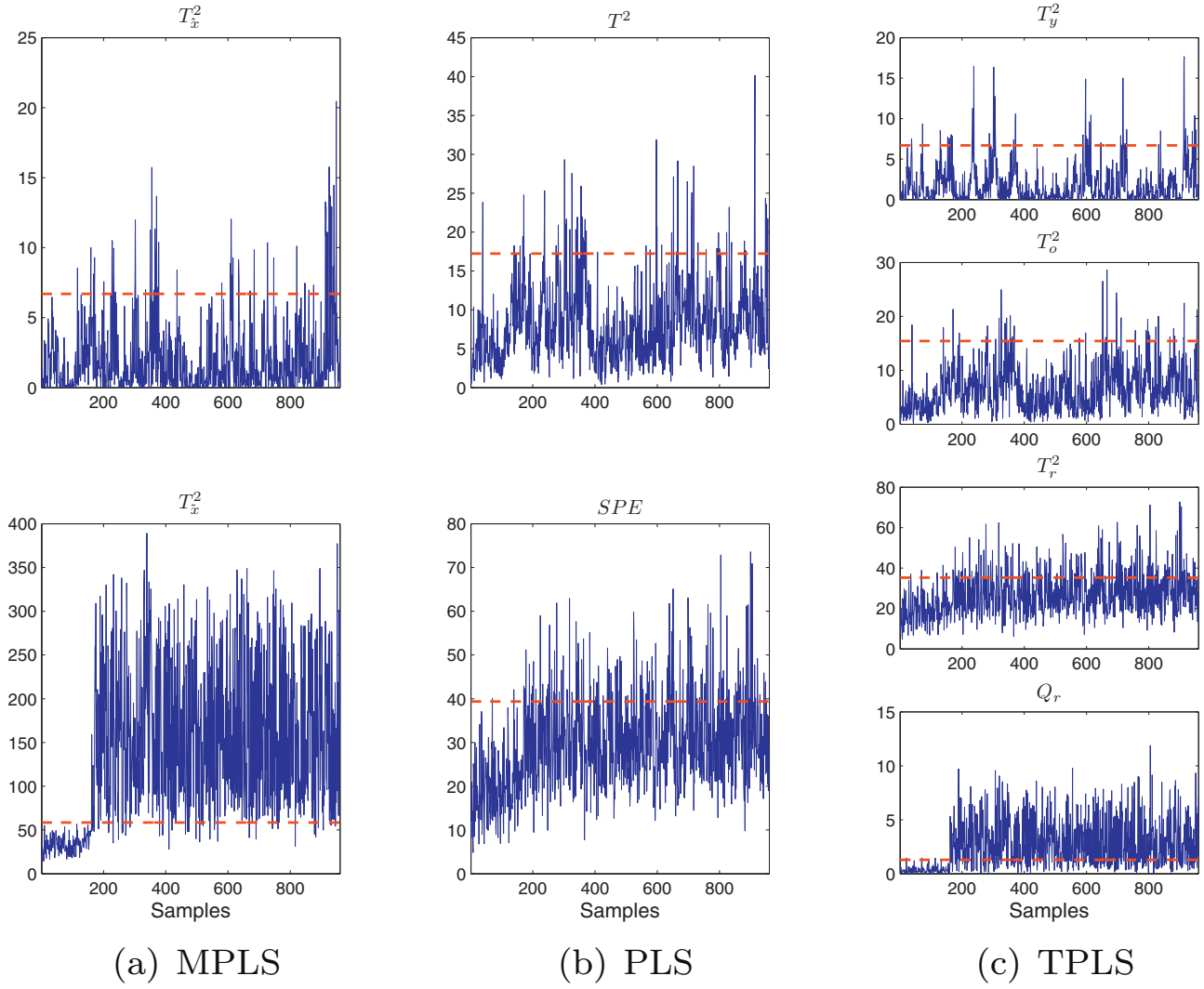


Fig. 7. Process monitoring using MPLS, PLS, TPLS in case of IDV(19).

Table 9
FDRs (%) based on the simulator given in [43].

Fault	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(7)	89.27	93.76	93.98	86.12	**94.78**	89.88	91.21	92.75	93.99
IDV(8)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(9)	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**
IDV(10)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(11)	94.70	**94.83**	94.64	94.64	94.64	99.68	99.65	99.64	99.64
IDV(12)	99.93	**99.95**	99.88	99.88	99.88	99.88	99.88	99.88	99.88
IDV(13)	99.72	**99.77**	99.63	99.63	99.63	99.63	99.63	99.63	99.63
IDV(1)	80.47	81.38	69.57	82.31	**83.35**	82.26	82.56	80.56	82.35
IDV(2)	46.21	51.33	**73.95**	59.94	66.03	50.68	56.56	63.40	65.88
IDV(4)	46.00	**60.22**	24.82	20.90	50.34	42.18	41.91	41.11	49.46
IDV(6)	56.63	63.71	87.77	80.08	87.90	54.83	75.28	84.90	**96.86**
IDV(17)	64.74	71.45	58.48	61.26	71.10	65.88	70.92	67.76	**83.11**
IDV(18)	21.09	30.66	30.69	19.63	33.42	21.32	31.68	26.52	**56.86**
IDV(20)	29.07	37.26	47.07	42.44	50.55	30.91	48.43	46.29	**61.26**
IDV(3)	1.28	1.50	1.23	1.46	3.34	1.93	**4.26**	1.79	2.92
IDV(5)	1.51	1.57	2.35	1.80	3.10	1.84	**4.20**	1.80	3.19
IDV(14)	1.41	1.39	2.50	2.25	3.22	1.83	**3.95**	1.58	3.50
IDV(15)	1.19	1.58	2.43	1.68	3.17	1.64	**3.74**	1.53	3.34
IDV(16)	1.43	1.69	3.03	1.69	3.13	1.61	**3.74**	1.56	2.59
IDV(19)	1.25	1.37	2.44	1.73	3.21	1.70	**3.92**	1.76	2.73

Table 10
FARs (%) based on the simulator given in [43].

Fault Free	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(0)	**1.26**	1.53	2.74	1.87	3.13	1.50	3.86	1.74	3.02

Table 11
FDRs (%) based on the simulator given in [43] with different design parameters.

Fault	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(8)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(9)	99.98	**99.99**	99.98	99.98	99.98	99.98	99.98	99.98	**99.99**
IDV(10)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IDV(11)	95.65	**95.75**	95.64	95.64	95.64	**95.75**	95.64	95.64	95.64
IDV(12)	99.89	**99.96**	99.88	99.88	99.88	99.88	99.88	99.88	99.88
IDV(13)	99.63	99.69	**99.75**	99.66	99.63	99.63	99.63	99.63	99.63
IDV(1)	75.70	75.30	72.91	77.55	**80.36**	76.53	76.97	76.42	72.55
IDV(2)	50.72	53.35	40.37	68.37	68.18	71.35	**77.63**	65.45	75.19
IDV(4)	55.12	70.03	49.03	32.22	54.22	43.03	42.39	43.94	**72.48**
IDV(6)	63.60	72.81	91.67	91.67	92.93	92.26	94.64	91.20	**99.01**
IDV(7)	87.35	89.55	87.49	84.74	88.92	86.80	87.14	87.12	**92.29**
IDV(17)	67.65	75.75	65.49	67.97	70.95	67.81	68.51	67.61	**87.32**
IDV(18)	24.95	36.25	18.70	21.69	30.78	24.86	25.93	24.35	**61.44**
IDV(20)	38.20	51.83	42.93	40.24	54.38	50.41	51.60	50.34	**71.76**
IDV(3)	1.53	2.17	2.72	2.17	3.25	1.53	2.90	1.48	**6.03**
IDV(5)	1.57	2.16	1.73	1.77	3.01	1.02	3.33	1.64	**4.07**
IDV(14)	1.67	2.29	1.61	2.22	3.05	1.46	2.90	1.51	**5.60**
IDV(15)	1.60	1.68	1.91	1.88	3.05	1.13	2.34	1.35	**3.70**
IDV(16)	1.84	1.88	2.30	1.84	3.20	1.27	2.63	1.46	**4.92**
IDV(19)	1.50	2.15	2.39	2.07	3.20	1.08	2.87	1.67	**4.31**

Table 12
FARs (%) based on the simulator given in [43] with different design parameters.

Fault Free	PCA	DPCA	ICA	MICA	FDA	PLS	TPLS	MPLS	SAP
IDV(0)	1.61	1.78	1.94	2.17	3.11	**1.45**	2.55	1.69	5.35

show better results than standard PCA but have not shown evident improvements over other methods. See Table 12.

5. Conclusions

In this paper, the basic data-driven PM–FD methods and their recent developments were firstly reviewed. The basic issues, including off-line design and on-line computation algorithms, original idea, basic assumption/condition and computation complexity were presented in detail. Then, all the discussed methods were implemented on an industrial benchmark of TE process to complete a detailed comparison study. As a result, we would like to point out that

- Standard PCA, which has not considered the autocorrelation of process variable, shows relatively lower FDRs compared with DPCA. Two variants of PLS, i.e. TPLS and MPLS, offer much better FDRs and more accurate fault diagnosis information compared with the standard approach. Although the ICA related methods involve complicated calculation, they only provide significant improvements compared to standard PCA approach. It is worth of further discussing whether the ICs, especially compared with the PCs, could bring additional advantages to the evaluation stage of PM–FD. Notice that the SAP provides superior FDRs in most cases due to its ability to deal with dynamic issue in the process with wide operating range of process variables.
- The design parameters in PCA, PLS and ICA related approaches will (considerably) influence the PM–FD performance. Although there are some criteria for parameter selection, it has not been analytically proved that which criterion offers best performance for PM–FD. Even for the same criterion, e.g. leave-N-out cross-validation to decide number of LVs, different results can be obtained according to different values of N. Hence, the methods like MPLS/TPLS and SAP, which do not have or are influenced little by such design parameters, have much more advantages in the application point of view.
- In practice, the large scale industrial plants are generally complex dynamic systems and the process measurements will not strictly follow Gaussian distribution as shown in TE process. On the other hand, it is also hard to give a physical explanation whether the non-Gaussian distributed process measurements can be described as a linear combination of the ICs. Although the process data cannot perfectly fulfill the basic assumptions in Table 1, most of the tested methods show their abilities for PM–FD in TE process even with non-Gaussian measurement noise. Especially, the method like SPA, which has higher FDRs, relatively lower computation cost and no special assumption on the process data, will receive more attentions both in practice application and in academic study.

Appendix A.

See Tables 13 and 14.

Table 13
Process variables.

Block name	Variable name	Number
Input feed	A feed (stream 1)	XMEAS(1)
	D feed (stream 2)	XMEAS(2)
	E feed (stream 3)	XMEAS(3)
	A and C feed	XMEAS(4)
Reactor	Reactor feed rate	XMEAS(6)
	Reactor pressure	XMEAS(7)
	Reactor level	XMEAS(8)
	Reactor temperature	XMEAS(9)
Separator	Separator temperature	XMEAS(11)
	Separator level	XMEAS(12)
	Separator pressure	XMEAS(13)
	Separator underflow	XMEAS(14)
Stripper	Stripper level	XMEAS(15)
	Stripper pressure	XMEAS(16)
	Stripper underflow	XMEAS(17)
	Stripper temperature	XMEAS(18)
	Stripper steam flow	XMEAS(19)
Miscellaneous	Recycle flow	XMEAS(5)
	Purge rate	XMEAS(10)
	Compressor work	XMEAS(20)
	Reactor water temperature	XMEAS(21)
	Separator water temperature	XMEAS(22)

Table 13 (*Continued*)

Block name	Variable name	Number
Reactor feed analysis	Component A	XMEAS(23)
	Component B	XMEAS(24)
	Component C	XMEAS(25)
	Component D	XMEAS(26)
	Component E	XMEAS(27)
	Component F	XMEAS(28)
Purge gas analysis	Component A	XMEAS(29)
	Component B	XMEAS(30)
	Component C	XMEAS(31)
	Component D	XMEAS(32)
	Component E	XMEAS(33)
	Component F	XMEAS(34)
	Component G	XMEAS(35)
	Component H	XMEAS(36)
Product analysis	Component D	XMEAS(37)
	Component E	XMEAS(38)
	Component F	XMEAS(39)
	Component G	XMEAS(40)
	Component H	XMEAS(41)

Table 14
Process manipulated variables.

Variable name	Number	Base value	Units
D feed flow	XMV(1)	63.053	$kg\,h^{-1}$
E feed flow	XMV(2)	53.980	$kg\,h^{-1}$
A feed flow	XMV(3)	24.644	ks cm h
A and C feed flow	XMV(4)	61.302	ks cm h
Compressor recycle valve	XMV(5)	22.210	%
Purge valve	XMV(6)	40.064	%
Separator pot liquid flow	XMV(7)	38.100	$m^3\,h^{-1}$
Stripper liquid product flow	XMV(8)	46.534	$m^3\,h^{-1}$
Stripper steam valve	XMV(9)	47.446	%
Reactor cooling water flow	XMV(10)	41.106	$m^3\,h^{-1}$
Condenser cooling water flow	XMV(11)	18.114	$m^3\,h^{-1}$
Agitator speed	XMV(12)	50.000	rpm

References

[1] M. Basseville, I. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice Hall, 1993.
[2] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, J. Schröder, Diagnosis and Fault-Tolerant Control, Springer-Verlag, Berlin, 2006.
[3] G. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification, The Annals of Mathematical Statistics 25 (2) (1954) 290–302.
[4] L. Chiang, M. Kotanchek, A. Kordon, Fault diagnosis based on fisher discriminant analysis and support vector machines, Computers and Chemical Engineering 28 (8) (2004) 1389–1401.
[5] L. Chiang, E. Russell, R. Braatz, Fault diagnosis in chemical processes using fisher discriminant analysis, discriminant partial least squares, and principal component analysis, Chemometrics and Intelligent Laboratory Systems 50 (2) (2000) 243–252.
[6] L. Chiang, E. Russell, R. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer-Verlag, London, 2001.
[7] B. Dayal, J. MacGregor, Improved PLS algorithms, Journal of Chemometrics 11 (1) (1997) 73–85.
[8] S. Ding, Model-based Fault Diagnosis Techniques, Springer-Verlag, Berlin, 2008.
[9] S. Ding, P. Zhang, E. Ding, P. Engel, W. Gui, A survey of the application of basic data-driven and model-based methods in process monitoring and fault diagnosis, in: Accepted for the 18th IFAC World Congress, Milano, Italy, 2011.
[10] S. Ding, P. Zhang, E. Ding, S. Yin, A. Naik, P. Deng, W. Gui, On the application of PCA technique to fault diagnosis, Tsinghua Science and Technology 15 (2) (2010) 138–144.
[11] S. Ding, P. Zhang, A. Naik, E. Ding, B. Huang, Subspace method aided data-driven design of fault detection and isolation systems, Journal of Process Control 19 (9) (2009) 1496–1510.
[12] S. Yin, S. Ding, A. Haghani, H. Hao, P. Zhang, Data-driven monitoring for stochastic systems and its application on batch process, International Journal of Systems Science (2012), http://dx.doi.org/10.1080/00207721.2012.659708.
[13] J. Downs, E. Fogel, A plant-wide industrial process control problem, Computers and Chemical Engineering 17 (1993) 245–255.
[14] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley–Interscience, New York, 2001.
[15] W. Favoreel, B.D. Moor, P.V. Overschee, Subspace state space system identification for industrial processes, Journal of Process Control 10 (2-3) (2000) 149–155.
[16] J. Gertler, Fault Detection and Diagnosis in Engineering Systems, Marcel Dekker Inc., New York, USA, 1998.
[17] M. Girolami, Self-Organising Neural Networks: Independent Component Analysis and Blind Source Separation, Springer-Verlag, London, 1999.
[18] Q. He, S. Qin, J. Wang, A new fault diagnosis method using fault directions in fisher discriminant analysis, AIChE Journal 51 (2) (2005) 555–571.
[19] I. Helland, On the structure of partial least squares regression, Communications in Statistics-Simulation and Computation 17 (1998) 581–607.
[20] A. Hoskuldsson, PLS regression methods, Journal of Chemometrics 2 (1998) 211–228.
[21] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Transactions on Neural Networks 10 (3) (1999) 626–634.
[22] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[23] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural networks 13 (4–5) (2000) 411–430.
[24] J. Jackson, A User's Guide to Principal Components, Wiley Interscience, New York, USA, 1991.
[25] J. Jackson, G. Mudholkar, Control procedures for residuals associated with principal component analysis, Technometrics 21 (3) (1979) 341–349.
[26] S.-L. Jämsä-Jounela, Future trends in process automation, Annual Reviews in Control 31 (2) (2007) 211–220.
[27] I. Jolliffe, Principal Component Analysis, Springer-Verlag, Berlin, 1986.
[28] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, Monitoring independent components for fault detection, AIChE Journal 49 (4) (2003) 969–976.
[29] A. Khan, J. Moyne, D. Tilbury, Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares, Journal of Process Control 18 (2008) 961–974.
[30] J. Kresta, J. MacGregor, T. Marlin, Multivariate statistical monitoring of process operating performance, Canadian Journal of Chemical Engineering 69 (1991) 35–47.
[31] W. Ku, R. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, Chemometrics and Intelligent Laboratory Systems 30 (1) (1995) 179–196.
[32] J.-M. Lee, S. Qin, I.-B. Lee, Fault detection and diagnosis based on modified independent component analysis, AIChE Journal 52 (10) (2006) 3501–3514.
[33] J.-M. Lee, C. Yoo, I.-B. Lee, Statistical process monitoring with independent component analysis, Journal of Process Control 14 (5) (2004) 467–485.
[34] T. Lee, Independent Component Analysis: Theory and Applications, Kluwer Academic, Boston, MA, 1998.
[35] G. Li, S. Qin, D. Zhou, Geometric properties of partial least squares for process monitoring, Automatica 46 (1) (2010) 204–210.
[36] E. Martin, A. Morris, Non-parametric confidence bounds for process performance monitoring charts, Journal of Process Control 6 (6) (1996) 349–358.
[37] G. McLACHLAN, Discriminant Analysis and Statistical Pattern Recognition, Wiley–Interscience, Hoboken, NJ, USA, 2004.
[38] P. Nomikos, J. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1) (1995) 41–59.
[39] P.V. Overschee, B.D. Moor, Subspace Identification for Linear Systems, Kluwer Academic Press, Dordrecht, 1996.
[40] R. Patton, P. Frank, R. Clark, Issues of Fault Diagnosis for Dynamic Systems, Springer-Verlag, Berlin, 2000.
[41] S. Qin, Statistical process monitoring: basics and beyond, Journal of Chemometrics 17 (2003) 480–502.
[42] S. Qin, Data-driven fault detection and diagnosis for complex industrial processes, in: Proceedings of the 7th IFAC Symposium on Fault Detection and Supervision and Safety of Technical Processes, Barcelona, Spain, 2009.
[43] N. Ricker, Decentralized control of the Tennessee Eastman challenge process, Journal of Process Control 6 (4) (1996) 205–221.
[44] E. Russell, L. Chiang, R. Braatz, Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes, Springer-Verlag, London, 2000.
[45] B. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, UK, 1986.
[46] N. Thornhill, A. Horch, Advances and new directions in plant-wide disturbance detection and diagnosis, Control Engineering Practice 15 (30) (2007) 1196–1206.
[47] N. Tracy, J. Young, R. Mason, Multivariate control charts for individual observations, Journal of Quality Technology 24 (2) (1992) 88–95.
[48] S. Valle, W. Li, S. Qin, Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods, Industrial and Engineering Chemistry Research 38 (11) (1999) 4389–4401.
[49] V. Venkatasubramanian, R. Rengaswamy, K. Yin, S. Kavuri, A review of process fault detection and diagnosis. Part III: process history based methods, Computers and Chemical Engineering 27 (3) (2003) 327–346.
[50] S. Yin, Data-driven design of fault diagnosis systems, VDI-Verlag, Düsseldorf, Germany, Fortschritt-Berichte, Reihe 8, Nr. 1206.
[51] M. Verhaegen, Identification of the deterministic part of MIMO state-space models given in innovations form from input–output data, Automatica 30 (1) (1994) 61–97.

[52] J. Wang, S. Qin, A new subspace identification approach based on principal component analysis, Journal of Process Control 12 (8) (2002) 841–855.
[53] B. Wise, N. Gallagher, The process chemometrics approach to process monitoring and fault detection, Journal of Process Control 6 (1996) 329–348.
[54] S. Wold, M. Sjostroma, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[55] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2) (2001) 109–130.
[56] S. Yin, S. Ding, P. Zhang, A. Haghani, A. Naik, Study on modifications of PLS approach for process monitoring, in: Accepted for the 18th IFAC World Congress, Milano, Italy, 2011.
[57] Y. Zhang, Y. Zhang, Fault detection of non-Gaussian processes based on modified independent component analysis, Chemical Engineering Science 65 (16) (2010) 4630–4639.
[58] D. Zhou, G. Li, S. Qin, Total projection to latent structures for process monitoring, AIChE Journal 56 (1) (2010) 168–178.