University of Reading

Continual learning-based probabilistic slow feature analysis for monitoring multimode nonstationary processes

Article

Accepted Version

Zhang, J., Zhou, D., Chen, M. and Hong, X. ORCID: https://orcid.org/0000-0002-6832-2298 (2022) Continual learning-based probabilistic slow feature analysis for monitoring multimode nonstationary processes. IEEE Transactions on Automation Science and Engineering, 21 (1). pp. 733-745. ISSN 1558-3783 doi: 10.1109/TASE.2022.3219125 Available at https://centaur.reading.ac.uk/108593/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1109/TASE.2022.3219125

Publisher: IEEE

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Continual learning-based probabilistic slow feature analysis for monitoring multimode nonstationary processes

Jingxin Zhang, Donghua Zhou, *Fellow, IEEE*, Maoyin Chen, *Member, IEEE*, and Xia Hong, *Senior Member, IEEE*

***Abstract*—A novel continual learning-based probabilistic slow feature analysis algorithm is introduced for monitoring multimode nonstationary processes. Multimode slow features are extracted and an elastic weight consolidation (EWC) is adopted for sequential modes. EWC was originally introduced in the setting of machine learning of sequential multi-tasks with the aim of avoiding catastrophic forgetting issue, which equally poses as a major challenge in multimode nonstationary process monitoring. When a new mode arrives, a small set of data are collected for continual learning by the proposed algorithm. A regularization term is introduced to prevent new data from significantly interfering with the learned knowledge, where the parameter importance measures are estimated. The proposed method is referred to as PSFA–EWC, which is updated continually and is capable of achieving excellent performance. PSFA–EWC furnishes backward and forward transfer ability by a single model. The significant features of previous modes are retained while consolidating new information, which may contribute to learning new relevant modes. The effectiveness of the proposed method is demonstrated via a continuous stirred tank heater and a practical coal pulverizing system.**

***Note to Practitioners*—Since industrial systems operate in varying modes and data are nonstationary within each mode, multimode nonstationary process monitoring is increasingly important. Traditional multimode monitoring methods generally need complete data from all possible modes and may need to be retrained from scratch when a new mode arrives, which require expensive computation and storage resources. Besides, it is difficult to distinguish real faults from normal variations in multimode nonstationary processes. This paper proposes a novel continual learning-based probabilistic slow feature analysis, where elastic weight consolidation is employed to consolidate the previously learned knowledge while extracting multimode slow features. The monitoring model is updated sequentially and provides backward as well as forward transfer learning ability for successive modes. It is able to separate real faults from normal dynamics, which is beneficial to identifying a new mode for multimode nonstationary processes. In addition, the proposed approach delivers excellent model interpretability and deals with missing data as well as uncertainty. In industrial applications, such as power plants and intelligent manufacturing processes, the proposed method can provide excellent monitoring performance.**

***Index Terms*—Multimode nonstationary processes, probabilistic slow feature analysis (PSFA), elastic weight consolidation (EWC), continual learning ability**

I. Introduction

Data-driven process monitoring is vitally important for ensuring safety and reliability of modern industrial processes [1]–[4]. Nonstationary process monitoring methods have been extensively studied [5]–[8]. Slow feature analysis (SFA), which is effective in extracting invariant slow features from fast changing sensing data [9], has been widely extended to process monitoring. SFA could establish a comprehensive operating status, where the nominal operating deviations and real faults may be distinguished in the closed-loop systems [10]–[13]. Recursive SFA (RSFA) [11] and recursive exponential SFA [12] were developed, and the associated parameters were updated for adaptive monitoring. Sufficient samples had been required to establish the initial model when a new mode was identified. Probabilistic SFA (PSFA) was proposed as a probabilistic framework with the advantage of effectively handling process noise and uncertainties, where measurement noise was modeled and missing data could be settled conveniently [13].

Most industrial systems operate in multiple conditions due to equipment maintenance, market demands, changing of raw materials, etc. Multimode nonstationary process monitoring methods have been investigated, which could be sorted into two categories [14], namely, single-model and multiple-model methods. Single-model methods transform the multimode data to unimodal distribution [15] or establish adaptive models [11], [16]. Local neighborhood standardization can normalize data into a single distribution and popular methods for one mode could be applied [15]. However, the effectiveness may be influenced by the matching degree of training and testing data. Although prior knowledge is not required, these algorithms are effective for slow changing features and may fail to track the dramatic variations on the entire dataset [11], [15], [16].

The mainstream approaches of multimode monitoring are based on multiple-model schemes, where the modes are identified and local models are built within each mode [17]–[19]. Mixture of canonical variate analysis (MCVA) was explored

This work was supported by National Natural Science Foundation of China [grant numbers 62033008, 61873143] and Taishan Scholar Project of Shandong Province of China. (Corresponding authors: Donghua Zhou; Maoyin Chen)

Jingxin Zhang is with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: zjx18@tsinghua.org.cn).

Donghua Zhou is with College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266000, China and also with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: zdh@mail.tsinghua.edu.cn).

Maoyin Chen is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: mychen@tsinghua.edu.cn).

Xia Hong is with Department of Computer Science, School of Mathematical, Physical and Computational Sciences, University of Reading, RG6 6AY, U.K.

for multimode nonstationary processes [18]. Besides, autoregressive dynamic latent variable was extended to multimode dynamic processes through a switching technique [20], where autocorrelations and cross-correlations were extracted based on a high-order Bayesian network model. In [21], a novel nonstationary discrete convolution kernel was proposed to deal with multimodality and nonlinear behavior, which aimed to overcome the limitation of radial basis function in multimode processes. Improved mixture of probabilistic principal component analysis (IMPPCA) could be utilized for multimode processes [19], where the model parameters and the mode identification were jointly optimized. However, the number of modes is a priori and data from all possible modes are required before learning, which is infeasible and time-consuming [14]. When novel modes appear, sufficient data should be collected and new local models are relearned correspondingly. The model is only effective for the learned modes, but may be difficult to deliver excellent performance for similar modes [18], [19]. Besides, multiple-model schemes may be redundant and difficult to identify modes accurately [22]. The model's capacity and storage costs increase significantly with the emergence of modes.

Recently, the emergent research area of continual learning has received much attention [23]–[28]. One long-standing challenge to be addressed is catastrophic forgetting issue, namely, learning a model with new information would influence the previously learned knowledge [23]. Continual learning is concerned with continual adaptation of the model to the changing tasks by acquiring new information while preserving the learned knowledge. While there are diverse techniques on continual learning ranging from regularization [23] to dynamic architectures [26] to manipulating data memory replay [24], the majority benchmarking applications in the literatures appear to focus on the image processing and generally require the class labels [23]–[27], [29]. Nevertheless, the concept of continual learning extends to lifelong machine learning [30] as well as poses open problems to related areas of machine learning such as transfer learning [31], etc (The readers are referred to [27] and references within). Of particular interest here is its integration with domain-specific learning such as autonomous agents [27] and conditioning monitoring [32]. One of the continual learning paradigms is called elastic weight consolidation (EWC) [23], in which it is analyzed that when a full data set of multiple tasks are decomposed based on a sequence of incoming tasks, the model parameters can be adjusted accordingly based on data from a new task, without sacrificing performance for any previously learned tasks. EWC was interpreted from Bayesian theory, thus providing excellent model interpretability [23].

Similarly, in the context of multimode process monitoring, new modes would often appear continuously and different modes may share similar significant features [22]. In practical applications, it is often intractable to collect data from all modes. Zhang *et al.* applied continual learning into multimode process monitoring [32], where EWC was employed to settle the catastrophic forgetting of principal component analysis (PCA), referred to as PCA–EWC. However, data are assumed to be stationary in each mode and a mode is identified by statistical characteristics of data, which makes it ineffective for multimode nonstationary processes, as well as difficult to distinguish the operating deviations and dynamic anomalies. Furthermore, a modified dynamic PCA with continual learning ability was presented for multimode dynamic processes and the mode identification was a priori [28], where modified synaptic intelligence was proposed and the parameter importance was measured by the sensitivity of each parameter to the loss. This method is free from the constraint that data should obey Gaussian distribution, but the importance may be influenced by the initial setting of the optimization issue. Sometimes, it may be intractable to accurately identify the mode switching only by prior knowledge.

Against this background, this paper considers a novel PSFA approach with continual learning ability, which is regarded as underlying multimode nonstationary processes for the observed sequential data. Moreover, the proposed algorithm would be best to distinguish real faults and normal operating derivations. When a new mode is identified by PSFA and limited prior knowledge, a small set of data are collected before learning. A quadratic penalty term is introduced to avoid the dramatic changes of mode-relevant parameters when a new mode is trained, where EWC is adopted to estimate the PSFA model parameter importance. PSFA assumes that the noise follows multivariate Gaussian distribution in each mode, which makes it possible to estimate parameter importance by EWC. The proposed method is referred to as PSFA–EWC. Since EWC can be interpreted from the perspective of Bayesian theory, the proposed method furnishes excellent model interpretability and solid theoretical foundation.

The contributions are summarized as follows:

a) PSFA with continual learning ability is firstly investigated for nonstationary processes, where data from multiple modes are collected in a sequential manner. The mode is identified by the statistics and limited prior knowledge, and the model is updated based on limited new data when a new mode arrives. PSFA–EWC provides excellent interpretability, and deals with missing data, measurement noise and uncertainty.
b) Compared with traditional multimode process monitoring methods, PSFA–EWC extracts new information and consolidates the previously learned knowledge simultaneously, which may aid the learning of future relevant modes and also be beneficial to monitoring the previously learned modes. Thus, PSFA–EWC furnishes the forward and backward transfer ability.
c) Compared with PCA–EWC [32], dynamic and static features are extracted and three monitoring statistics are designed, which can distinguish partial normal variations and real faults. Besides, the importance is calculated by the covariance of the gradient of the model's log likelihood function with respect to the local optimum, instead of the expectation of second-order derivative, which is more suitable for large-scale industrial systems.

The rest of this paper is organized below. Section II reviews PSFA succinctly and outlines the basic idea of our proposed approach. The technical core of PSFA–EWC is detailed in

Section III. The monitoring procedure and comparative experiments are designed in Section IV. The effectiveness of PSFA–EWC is illustrated by a continuous stirred tank heater (CSTH) and a practical coal pulverizing system in Section V. The conclusion is given in Section VI.

II. Preliminaries and Problem Statement

For ease of exposition, we start with introducing the PSFA for a single mode, since it serves as basic ingredient of our proposed multimode PSFA. Then the basic idea of EWC as well as how to extend EWC to multimode PSFA is outlined.

A. PSFA for a single mode

PSFA aims to identify the slowest varying latent features from a sequence of time-varying observations $\boldsymbol{x}_t \in R^m$, $t = 1, 2, \ldots, T$, which can be represented/generated via a state-space model with a first-order Markov chain architecture [33].

$$\begin{aligned} \boldsymbol{x}_t =& \boldsymbol{V}\boldsymbol{y}_t + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_x) \\ \boldsymbol{y}_t =& \boldsymbol{\Lambda}\boldsymbol{y}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{y}_1 =& \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_1) \end{aligned} \tag{1}$$

where the low dimensional latent variable $\boldsymbol{y}_t \in R^p$, $p < m$. $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$, with the constraint $\boldsymbol{\Lambda}^2 + \boldsymbol{\Sigma} = \boldsymbol{I}$ to ensure the covariance matrix be the unit matrix $\boldsymbol{I}$. The emission matrix is $\boldsymbol{V} \in R^{m\times q}$ and measurement noise variance is $\boldsymbol{\Sigma}_x = diag\left(\sigma_1^2, \cdots, \sigma_m^2\right)$.

For a single mode, the observed data and latent slow features sequences are denoted as $\boldsymbol{X}_s = \{\boldsymbol{x}_t\} \in R^{m\times T}$ and $\boldsymbol{Y}_s = \{\boldsymbol{y}_t\} \in R^{p\times T}$, respectively. T is the number of samples and the estimation of p has been discussed in [13].

The joint distribution is given as [34]

$$P(\boldsymbol{X}_s|\boldsymbol{Y}_s) = P(\boldsymbol{y}_1)\prod_{t=2}^{T} P\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}\right)\prod_{t=1}^{T} P(\boldsymbol{x}_t|\boldsymbol{y}_t) \tag{2}$$

Let $\theta_x = \{\boldsymbol{V}, \boldsymbol{\Sigma}_x\}$, $\theta_y = \{\boldsymbol{\Sigma}_1, \boldsymbol{\Lambda}\}$. The objective of PSFA is to estimate parameters $\theta = \{\theta_x, \theta_y\}$ by maximizing the complete log likelihood function:

$$\begin{aligned} \log\ P(\boldsymbol{X}_s, \boldsymbol{Y}_s|\theta) =& \sum_{t=1}^{T} \log\ P(\boldsymbol{x}_t|\boldsymbol{y}_t, \theta_x) + \log\ P(\boldsymbol{y}_1|\boldsymbol{\Sigma}_1) \\ &+ \sum_{t=2}^{T} \log\ P\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{\Lambda}\right) \end{aligned} \tag{3}$$

The optimal parameter θ is optimized by maximizing (3) using expectation maximization (EM) algorithm [35].

B. Problem statement

The problem is interpreted for multimode nonstationary processes and then the key objective is summarized. Consider also based on PSFA model (1), in the multimode scenario where data stream are generated as incoming new modes $\mathcal{M}_K$, $K = 1, 2, \ldots$ one at time. For each mode $\mathcal{M}_K$, normal data $\boldsymbol{X}_K \in R^{m\times T_K}$ are collected, where T_K is the number of samples. Correspondingly it is assumed that $\boldsymbol{Y}_K \in R^{p\times T_K}$ need to be extracted from the Kth mode.

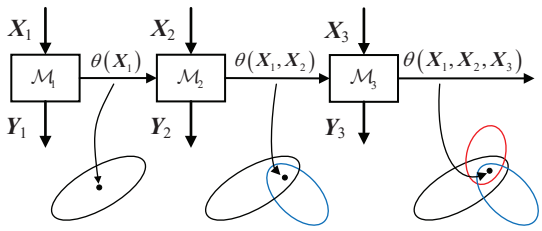


Fig. 1. An illustration of the proposed PSFA–EWC with continual learning ability for three consecutive modes $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$.

Denote the total observed data and its latent slow features as $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\}$, $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots\}$.

EWC initially considers Bayesian rule for the sequential learning process in which the most probable parameters should be found by maximizing the conditional probability [23]

$$\log\ P(\theta|\boldsymbol{X}, \boldsymbol{Y}) = \log\ P(\boldsymbol{X}, \boldsymbol{Y}|\theta) + \log\ P(\theta) - \log\ P(\boldsymbol{X}, \boldsymbol{Y}) \tag{4}$$

where $P(\theta)$ is prior probability and $P(\boldsymbol{X}, \boldsymbol{Y}|\theta)$ is the data probability. For illustration only the first two successive independent modes $\mathcal{M}_1$ and $\mathcal{M}_2$ are initially considered. Then, (4) can be reformulated as [23]:

$$\begin{aligned} \log\ P(\theta|\boldsymbol{X}, \boldsymbol{Y}) =& \log\ P(\boldsymbol{X}_2, \boldsymbol{Y}_2|\theta) + \log\ P(\theta|\boldsymbol{X}_1, \boldsymbol{Y}_1) \\ &+ don't\ care\ terms \end{aligned} \tag{5}$$

where $P(\theta|\boldsymbol{X}, \boldsymbol{Y})$ is the posterior probability of the parameter given the entire dataset. $P(\boldsymbol{X}_2, \boldsymbol{Y}_2|\theta)$ represents the loss function for mode $\mathcal{M}_2$. Posterior distribution $P(\theta|\boldsymbol{X}_1, \boldsymbol{Y}_1)$ can reflect all information of mode $\mathcal{M}_1$ [23]. This equation reflects the key idea of EWC in continual learning framework of updating system parameters based on a composite cost function that is dependent on current parameters learned from previous data and new incoming data by using posterior distribution $P(\theta|\boldsymbol{X}_1, \boldsymbol{Y}_1)$ which acts as a constraint in future objective, so that the learned knowledge will not be forgotten.

This is the first time that continual learning-based PSFA is proposed for monitoring where new optimization procedures of PSFA–EWC will be introduced in Section III, as depicted in Fig. 1 for three modes. The multimode slow features for each mode are extracted, while the parameter θ is continually updated using only data of a new mode, while maintaining performance of all old modes. Black, blue and red circles represent the optimal parameter regions that the log likelihoods for modes $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are maximized, respectively. This process can be generalized to $K > 3$ modes, with

$$\begin{aligned} \log\ P(\theta|\boldsymbol{X}, \boldsymbol{Y}) =& \log\ P(\boldsymbol{X}_K, \boldsymbol{Y}_K|\theta) + \log\ P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right) \\ &+ don't\ care\ terms \end{aligned} \tag{6}$$

where $P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right) \triangleq P(\theta|\boldsymbol{X}_1, ..., \boldsymbol{X}_{K-1}, \boldsymbol{Y}_1, ..., \boldsymbol{Y}_{K-1})$.

The first term in (6) is complete likelihood for Kth mode. The second term in (6) is parameter estimate that reflects information from all previous modes, thus can be interpreted as log prior probability of parameter for Kth mode. Since it is assumed that data from all previous modes will not be accessible to obtain $P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right)$ exactly, it is found by recursive approximation as detailed in Section III-A.

III. The proposed PSFA–EWC algorithm

A. Recursive Laplace approximation of $P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right)$

Consider the multimode PSFA process where data are collected sequentially with mode index $K = 1, 2, 3, ...$ For the sake of notational simplicity, it is assumed in the sequel that the data $\boldsymbol{x}_t$ and corresponding slow features $\boldsymbol{y}_t$ start from $t = 1$ at the beginning and end at $t = T_K$ of the Kth mode. The proposed PSFA–EWC algorithm starts with solving an initial single mode model as $K = 1$. An optimal parameter, denoted as $\theta^*_{\mathcal{M}_1}$, has been obtained from the first mode $\mathcal{M}_1$ based on solving (3). For later modes ($K \geq 2$), the monitoring model is updated recursively based on the data from Kth mode and the current monitoring model before K, where EM is employed [35] to solve the optimization problem of maximizing $J(\theta)$ in Section III-B.

Consider initially two modes $K = 2$, $\log P(\theta|\mathcal{M}_1)$ in (6) is approximated by the Laplace approximation [23] as

$$\log P(\theta|\mathcal{M}_1) \approx -\frac{1}{2}(\theta - \theta^*_{\mathcal{M}_1})^T \left(T_1 \boldsymbol{F}(\theta^*_{\mathcal{M}_1}) + \lambda_{prior}\boldsymbol{I}\right) \cdot (\theta - \theta^*_{\mathcal{M}_1}) + constant$$

where $\boldsymbol{F}(\theta^*_{\mathcal{M}_1})$ is Fisher information matrix (FIM) and computed by (27) in Appendix A. $\lambda_{prior}\boldsymbol{I}$ is the Gaussian prior precision matrix for mode $\mathcal{M}_1$. The sample size T_1 may have non-negligible influence on the approximation, which would be replaced by a mode-specific hyperparameter $\eta_1 > 0$ to enhance the approximation quality [36], namely,

$$\log P(\theta|\mathcal{M}_1) = -\frac{1}{2}(\theta - \theta^*_{\mathcal{M}_1})^T \boldsymbol{\Omega}_{\mathcal{M}_1}(\theta - \theta^*_{\mathcal{M}_1}) + constant$$

where $\boldsymbol{\Omega}_{\mathcal{M}_1} = \eta_1 \boldsymbol{F}(\theta^*_{\mathcal{M}_1}) + \lambda_{prior}\boldsymbol{I}$.

When the Kth mode $\mathcal{M}_K$ arrives ($K \geq 3$), we approximate $\log P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right)$ by recursive Laplace approximation as

$$\log P(\theta|\mathcal{M}_{i=1}^{K-1}) \approx -\frac{1}{2}(\theta - \theta^*_{\mathcal{M}_{K-1}})^T \boldsymbol{\Omega}_{\mathcal{M}_{K-1}}(\theta - \theta^*_{\mathcal{M}_{K-1}}) + constant$$

where

$$\boldsymbol{\Omega}_{\mathcal{M}_{K-1}} = \boldsymbol{\Omega}_{\mathcal{M}_{K-2}} + \eta_{K-1}\boldsymbol{F}_{\mathcal{M}_{K-1}}, \quad K \geq 3 \tag{7}$$

$\boldsymbol{F}_{\mathcal{M}_{K-1}}$ is FIM of mode $\mathcal{M}_{K-1}$ and η_{K-1} is a hyperparameter. $\log\ P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right)$ is approximated by a quadratic term centered at current optimum, with $\boldsymbol{\Omega}_{\mathcal{M}_{K-1}}$ acting as an importance measure regulating data from $K-1$ modes.

B. PSFA–EWC algorithm

Consider the objective of PSFA–EWC of maximizing

$$J(\theta) = \log\ P(\boldsymbol{X}_K, \boldsymbol{Y}_K|\theta) + \log P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right) \tag{8}$$

subject to PSFA model (1). Recall (3), the log-likelihood function for the current mode $\mathcal{M}_K$ is represented by

$$\log\ P(\boldsymbol{X}_K, \boldsymbol{Y}_K|\theta) = \sum_{t=1}^{T_K} \log\ P(\boldsymbol{x}_t|\boldsymbol{y}_t, \theta_x) + \log\ P(\boldsymbol{y}_1|\boldsymbol{\Sigma}_1) + \sum_{t=2}^{T_K} \log\ P\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{\Lambda}\right) \tag{9}$$

The regularization term is designed as

$$\log P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right) \approx -\gamma_{1,K}\left\|\boldsymbol{V} - \boldsymbol{V}_{\mathcal{M}_{K-1}}\right\|^2_{\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}} - \gamma_{2,K}\sum_{i=1}^{p}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\left(\lambda_i - \lambda_{\mathcal{M}_{K-1},i}\right)^2 \tag{10}$$

where $\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}$ and $\Omega^{\lambda}_{\mathcal{M}_{K-1},i}$ measure the importance of $\boldsymbol{V}_{\mathcal{M}_{K-1}}$ and $\lambda_{\mathcal{M}_{K-1},i}$, $i = 1, \cdots, p$. $\lambda_{\mathcal{M}_{K-1},i}$ and $\Omega^{\lambda}_{\mathcal{M}_{K-1},i}$ are the ith elements of diagonal matrices $\boldsymbol{\Lambda}_{\mathcal{M}_K}$ and $\boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}}$, which are the optimal parameters of mode $\mathcal{M}_{K-1}$. $\gamma_{1,K}$ and $\gamma_{2,K}$ are user-defined hyperparameters. The setting $\gamma_{1,K}$ and $\gamma_{2,K}$ makes it flexible to adjust the weights of previous modes.

For the proposed PSFA–EWC, the total objective function of K modes can be formally described by

$$J(\theta) = \sum_{t=1}^{T_K} \log\ P(\boldsymbol{x}_t|\boldsymbol{y}_t, \theta_x) + \sum_{t=2}^{T_K} \log\ P\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{\Lambda}\right) + \log\ P(\boldsymbol{y}_1|\boldsymbol{\Sigma}_1) - \gamma_{1,K}\left\|\boldsymbol{V} - \boldsymbol{V}_{\mathcal{M}_{K-1}}\right\|^2_{\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}} - \gamma_{2,K}\sum_{i=1}^{p}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\left(\lambda_i - \lambda_{\mathcal{M}_{K-1},i}\right)^2 \tag{11}$$

subject to the PSFA model (1). Note that for $K > 2$, since the quadratic penalty is added, it slows down the changes to parameters with respect to the previous optimum values that are obtained in learned modes [24], [27]. In other words, the parameters that result in significant deterioration in performance of previous modes will be penalized, avoiding catastrophic forgetting problem.

When $K = 1$, $\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}} = \boldsymbol{0}$, $\boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}} = \boldsymbol{0}$. There is no need to provide $\boldsymbol{V}_{\mathcal{M}_{K-1}}$ and $\boldsymbol{\Lambda}_{\mathcal{M}_{K-1}}$, this means that the proposed PSFA–EWC algorithm has a unified formulation as a sequential single mode based on Kth mode data only, with current parameters used as quadratic penalty, which are updated via recursive Laplace approximation between each mode in Section III-A. The EM [35] is employed to optimize the parameter $\theta = \{\boldsymbol{V}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_1\}$ by solving (11).

*1) **E-step**:* Assume that θ is available, the E-step estimates three sufficient statistics, namely,

$$\mathbb{E}\left[\boldsymbol{y}_t|\boldsymbol{X}_K\right] = \hat{\boldsymbol{\mu}}_t \tag{12}$$

$$\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_{t-1}^T|\boldsymbol{X}_K\right] = \boldsymbol{J}_{t-1}\hat{\boldsymbol{U}}_t + \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_{t-1}^T \tag{13}$$

$$\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^T|\boldsymbol{X}_K\right] = \hat{\boldsymbol{U}}_t + \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_t^T \tag{14}$$

Detailed information has been summarized in Appendix B.

*2) **M-step**:* Assume that three sufficient statistics are fixed, the parameters are updated alternately.

Since $\boldsymbol{V}$ and $\boldsymbol{\Sigma}_x$ are contained in $P(\boldsymbol{x}_t, \boldsymbol{y}_t|\theta_x)$ and the regularization term $\gamma_{1,K}\left\|\boldsymbol{V} - \boldsymbol{V}_{\mathcal{M}_{K-1}}\right\|^2_{\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}}$, then

$$\{\boldsymbol{V}^{new}, \boldsymbol{\Sigma}_x^{new}\} = \arg\max_{\boldsymbol{V}, \boldsymbol{\Sigma}_x}\ J(\boldsymbol{V}, \boldsymbol{\Sigma}_x) \tag{15}$$

where

$$
\begin{aligned}
&J\left(\boldsymbol{V}, \boldsymbol{\Sigma}_x\right) \\
=&\sum_{t=1}^{T_K} \log\ P\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\theta_x\right) - \gamma_{1,K}\left\|\boldsymbol{V}-\boldsymbol{V}_{\mathcal{M}_{K-1}}\right\|^2_{\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}} \\
=&-\frac{T_K}{2}\log|\boldsymbol{\Sigma}_x| - \frac{1}{2}\sum_{t=1}^{T_K}\Big(tr\left(\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^T|\boldsymbol{X}_K\right]\right)\boldsymbol{V}^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V} \\
&+tr\left(\boldsymbol{x}_t\boldsymbol{x}_t^T\boldsymbol{\Sigma}_x^{-1}\right) - 2tr\left(\boldsymbol{x}_t^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\mathbb{E}\left[\boldsymbol{y}_t|\boldsymbol{X}_K\right]\right)\Big) \\
&-\gamma_{1,K}tr\left((\boldsymbol{V}-\boldsymbol{V}_{\mathcal{M}_{K-1}})^T\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}(\boldsymbol{V}-\boldsymbol{V}_{\mathcal{M}_{K-1}})\right)
\end{aligned}
$$

Let the derivative with respect to $\boldsymbol{V}$ be zero, then

$$
\begin{aligned}
&\sum_{t=1}^{T_K}\boldsymbol{x}_t\mathbb{E}\left[\boldsymbol{y}_t^T|\boldsymbol{X}_K\right] + \gamma_{1,K}\boldsymbol{\Sigma}_x\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}\boldsymbol{V}_{\mathcal{M}_{K-1}} \\
&=\boldsymbol{V}\sum_{t=1}^{T_K}\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^T|\boldsymbol{X}_K\right] + \gamma_{1,K}\boldsymbol{\Sigma}_x\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}\boldsymbol{V}
\end{aligned} \tag{16}
$$

This problem is actually the Sylvester equation and the solution is denoted as $\boldsymbol{V}^{new}$.

Taking the derivative about σ_i^2 and let it be zero, then

$$
\begin{aligned}
\left(\sigma_i^2\right)^{new} =&\frac{1}{T_K}\sum_{t=1}^{T_K}\Big\{\mathbb{E}\left[x_{t,i}^2\right] - 2\left(\boldsymbol{v}_{\cdot i}^T\right)^{new}\mathbb{E}\left[\boldsymbol{y}_t|\boldsymbol{X}_K\right]x_{t,i} \\
&+\left(\boldsymbol{v}_{\cdot i}^T\right)^{new}\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^T|\boldsymbol{X}_K\right]\left(\boldsymbol{v}_{\cdot i}\right)^{new}\Big\}
\end{aligned} \tag{17}
$$

where $\left(\boldsymbol{v}_{\cdot i}^T\right)^{new}$ is the ith row of matrix $\boldsymbol{V}^{new}$, $1 \le i \le m$, and $\boldsymbol{\Sigma}_x^{new} = diag\left(\left(\sigma_1^2\right)^{new}, \cdots, \left(\sigma_m^2\right)^{new}\right)$.

With regard to $\boldsymbol{\Sigma}_1$, it is only contained in $P\left(\boldsymbol{y}_1\right)$, thus

$$
\begin{aligned}
\boldsymbol{\Sigma}_1^{new} =&\arg\max_{\boldsymbol{\Sigma}_1}\ \mathbb{E}\left[\log P\left(\boldsymbol{y}_1|\boldsymbol{\Sigma}_1\right)\right] \\
=&\mathbb{E}\left[\boldsymbol{y}_1\boldsymbol{y}_1^T|\boldsymbol{X}_K\right]
\end{aligned} \tag{18}
$$

For $\boldsymbol{\Lambda} = diag\left(\lambda_1, \cdots, \lambda_p\right)$, $\boldsymbol{\Sigma} = \boldsymbol{I} - \boldsymbol{\Lambda}^2$. λ_i is contained in $\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\left(\lambda_i - \lambda_{\mathcal{M}_{K-1},i}\right)^2$ and $P\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{\Lambda}\right)$, thus

$$
\boldsymbol{\Lambda}^{new} = \arg\max_{\boldsymbol{\Lambda}}\ J\left(\boldsymbol{\Lambda}\right)
$$

where

$$
\begin{aligned}
&J\left(\boldsymbol{\Lambda}\right) \\
=&\sum_{t=2}^{T_K}\log\ P\left(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{\Lambda}\right) - \gamma_{2,K}\sum_{i=1}^{p}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\left(\lambda_i - \lambda_{\mathcal{M}_{K-1},i}\right)^2 \\
=&-\frac{1}{2}\sum_{t=2}^{T_K}\sum_{i=1}^{p}\Big[\log\left(1-\lambda_i^2\right) + \frac{1}{1-\lambda_i^2}\left(\mathbb{E}\left[y_{t,i}^2|\boldsymbol{X}_K\right]\right. \\
&\left.-2\lambda_i\mathbb{E}\left[y_{t,i}y_{t-1,i}|\boldsymbol{X}_K\right] + \lambda_i^2\mathbb{E}\left[y_{t-1,i}^2|\boldsymbol{X}_K\right]\right)\Big] \\
&-\gamma_{2,K}\sum_{i=1}^{p}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\left(\lambda_i - \lambda_{\mathcal{M}_{K-1},i}\right)^2
\end{aligned}
$$

Let the derivative with respect λ_i be zero, then

$$
a_{i5}\lambda_i^5 + a_{i4}\lambda_i^4 + a_{i3}\lambda_i^3 + a_{i2}\lambda_i^2 + a_{i1}\lambda_i + a_{i0} = 0 \tag{19}
$$

where the coefficients of (19) are derived as

$$
\begin{aligned}
a_{i5} =&2\gamma_{2,K}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}, \\
a_{i4} =&-2\gamma_{2,K}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\lambda_{\mathcal{M}_{K-1},i}, \\
a_{i3} =&T_K - 1 - 4\gamma_{2,K}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}, \\
a_{i2} =&4\gamma_{2,K}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\lambda_{\mathcal{M}_{K-1},i} - \sum_{t=2}^{T_K}\mathbb{E}\left[y_{t,i}y_{t-1,i}|\boldsymbol{X}_K\right], \\
a_{i1} =&2\gamma_{2,K}\Omega^{\lambda}_{\mathcal{M}_{K-1},i} + \sum_{t=2}^{T_K}\left(\mathbb{E}\left[y_{t-1,i}^2|\boldsymbol{X}_K\right] + \mathbb{E}\left[y_{t,i}^2|\boldsymbol{X}_K\right] - 1\right), \\
a_{i0} =&-2\gamma_{2,K}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\lambda_{\mathcal{M}_{K-1},i} - \sum_{t=2}^{T_K}\mathbb{E}\left[y_{t,i}y_{t-1,i}|\boldsymbol{X}_K\right]
\end{aligned}
$$

Thus, the updated λ_i^{new} could be calculated numerically as the root of (19) within the range $[0, 1)$, and $\boldsymbol{\Lambda}^{new} = diag\left(\lambda_1^{new}, \cdots, \lambda_p^{new}\right)$.

The learning procedure of PSFA–EWC is summarized in Algorithm 1. The transformation and emission matrices are denoted as $\boldsymbol{\Lambda}_{\mathcal{M}_K}$ and $\boldsymbol{V}_{\mathcal{M}_K}$, respectively. Since noise information about $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_x$ is only effective for the current mode, the subscript $\mathcal{M}_K$ is neglected.

After the mode $\mathcal{M}_K$ has been learned, the importance measures specific to PSFA are updated and ready as $(K+1)$th mode.

$$
\boldsymbol{\Omega}^V_{\mathcal{M}_K} = \boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}} + \eta_K^V\boldsymbol{F}^V_{\mathcal{M}_K} \tag{20}
$$

$$
\boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_K} = \boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}} + \eta_K^{\Lambda}\boldsymbol{F}^{\Lambda}_{\mathcal{M}_K} \tag{21}
$$

where $\boldsymbol{F}^V_{\mathcal{M}_K}$ and $\boldsymbol{F}^{\Lambda}_{\mathcal{M}_K}$ are calculated by (29) and (31). η_K^V and η_K^{Λ} are mode-specific hyperparamaters, which are optimized by hyperparameter search [23] and fine-tuned by prior knowledge, and may play an important role in accurate estimate of probability with sequential modes. Then we illustrate the difference $\gamma_{1,K}$ and $\gamma_{2,K}$ to η^{Λ}_{K-1} and η^V_{K-1}. Combined with the importance of current mode $\mathcal{M}_K$, the setting $\gamma_{1,K}$ and $\gamma_{2,K}$ is beneficial to assigning the importance of all previous $K-1$ modes again. η^{Λ}_{K-1} and η^V_{K-1} focus on the importance of the mode $\mathcal{M}_{K-1}$, which allow users to obtain models with more focus on a particular mode.

IV. Monitoring procedure and experiment design

Analogous to traditional PSFA [13], three monitoring statistics are designed to provide a comprehensive operating status. Then, several representative methods are adopted as comparisons to illustrate the superiorities of PSFA–EWC algorithm.

A. Monitoring procedure

In this paper, the Hotelling's T^2 and SPE statistics are used to reflect the steady variations, and S^2 is calculated to evaluate the temporal dynamics [13].

According to Kalman filter equation,

$$
\boldsymbol{y}_t = \boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{y}_{t-1} + \boldsymbol{K}\left[\boldsymbol{x}_t - \boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{y}_{t-1}\right] \tag{22}
$$

After the training phase, $\boldsymbol{K}_t$ would converge to a steady matrix $\boldsymbol{K}$. Then, T^2 statistic is defined as

$$
T^2 = \boldsymbol{y}_t^T\boldsymbol{y}_t \tag{23}
$$

To design SPE, the bias between the true value and one-step prediction is calculated at t instant. At $(t-1)$ instant, the inferred slow features follow Gaussian distribution, namely,

$$P\left(\boldsymbol{y}_{t-1}|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{t-1}\right)\sim N\left(\boldsymbol{\mu}_{t-1},\boldsymbol{P}_{t-1}\right)$$

Then, the conditional distribution of $\boldsymbol{y}_t$ is described as

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{t-1}\right)\sim N\left(\boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{\mu}_{t-1},\boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{P}_{t-1}\boldsymbol{\Lambda}_{\mathcal{M}_K}^T+\boldsymbol{\Sigma}\right)$$

Similarly,

$$P\left(\boldsymbol{x}_t|\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{t-1}\right)\sim N\left(\boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{\mu}_{t-1},\boldsymbol{\Phi}_t\right)$$

where $\boldsymbol{\Phi}_t=\boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{P}_{t-1}\boldsymbol{\Lambda}_{\mathcal{M}_K}^T\boldsymbol{V}_{\mathcal{M}_K}^T+\boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{\Sigma}\boldsymbol{V}_{\mathcal{M}_K}^T+\boldsymbol{\Sigma}_x$. The prediction error follows Gaussian distribution, namely

$$\boldsymbol{\varepsilon}_t=\boldsymbol{x}_t-\boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{\Lambda}_{\mathcal{M}_K}\boldsymbol{\mu}_{t-1}\sim N\left(\boldsymbol{0},\boldsymbol{\Phi}_t\right)\tag{24}$$

After the training phase, $\boldsymbol{\Phi}_t$ converges to $\boldsymbol{\Phi}$. The SPE statistic is calculated by

$$SPE=\boldsymbol{\varepsilon}_t^T\boldsymbol{\Phi}^{-1}\boldsymbol{\varepsilon}_t\tag{25}$$

S^2 statistic is designed to reflect the temporal dynamics, which is beneficial to distinguishing the normal operating variations and dynamics anomalies [10], [13].

$$S^2=\dot{\boldsymbol{y}}_t^T\boldsymbol{\Xi}^{-1}\dot{\boldsymbol{y}}_t\tag{26}$$

where $\dot{\boldsymbol{y}}_t=\boldsymbol{y}_t-\boldsymbol{y}_{t-1}$, $\boldsymbol{\Xi}=\mathbb{E}\left\{\dot{\boldsymbol{y}}_t\dot{\boldsymbol{y}}_t^T\right\}$ is the covariance matrix and analytically calculated as $\boldsymbol{\Xi}=2\left(\boldsymbol{I}_p-\boldsymbol{\Lambda}_{\mathcal{M}_K}\right)$ [13].

The thresholds of three statistics are calculated by kernel density estimation (KDE) [19], and denoted as J_{th,T^2}, $J_{th,SPE}$ and J_{th,S^2}. The monitoring rule is summarized below:

1) All statistics are within thresholds, the process is normal;
2) If T^2 or SPE is over its threshold, while S^2 is below its threshold, the dynamic law remains unchanged and the static variations occur. This may be caused by step faults or normal drifts [10], [12], which has been explained in the supplementary material. Besides, this situation should be confirmed and distinguished further based on data and limited prior knowledge. When a new mode occurs, a small set of new data are collected to update the PSFA–EWC model. The process is monitored by S^2 statistic before the updating procedure;
3) If S^2 is over threshold, the dynamic behaviors are unusual and the system is out of control. A fault occurs and the alarm would be triggered.

Algorithm 1 Off-line training procedure of PSFA–EWC

Input: $\tilde{\boldsymbol{X}}_K$, $\boldsymbol{V}_{\mathcal{M}_{K-1}}$, $\boldsymbol{\Lambda}_{\mathcal{M}_{K-1}}$, $\boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}$, $\boldsymbol{\Omega}^\Lambda_{\mathcal{M}_{K-1}}$

Output: $\boldsymbol{\mu}_K$, $\boldsymbol{\Sigma}_K$, $\boldsymbol{V}_{\mathcal{M}_K}$, $\boldsymbol{\Lambda}_{\mathcal{M}_K}$, $\boldsymbol{\Omega}^V_{\mathcal{M}_K}$, $\boldsymbol{\Omega}^\Lambda_{\mathcal{M}_K}$, $\boldsymbol{K}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Xi}$, J_{th,T^2}, $J_{th,SPE}$ and J_{th,S^2}

1: For the mode $\mathcal{M}_K$, collect normal data $\tilde{\boldsymbol{X}}_K$, calculate the mean $\boldsymbol{\mu}_K$ and standard variance $\boldsymbol{\Sigma}_K$. Normalize data and the scaled data are denoted as $\boldsymbol{X}_K$;
2: Initialize parameters $\boldsymbol{V}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}$;
3: **While** the issue (11) is not converged **do**
   a) Calculate three sufficient statistics by (12)–(14);
   b) Update the parameters by (16)–(19);
4: The optimal emission and transition matrices are denoted as $\boldsymbol{V}_{\mathcal{M}_K}$ and $\boldsymbol{\Lambda}_{\mathcal{M}_K}$, respectively;
5: Calculate the FIMs $\boldsymbol{F}^V_{\mathcal{M}_K}$ by (29) and $\boldsymbol{F}^\Lambda_{\mathcal{M}_K}$ by (31). Then, update the importance measures $\boldsymbol{\Omega}^V_{\mathcal{M}_K}$ and $\boldsymbol{\Omega}^\Lambda_{\mathcal{M}_K}$ by (20)–(21);
6: The final Kalman matrix is denoted as $\boldsymbol{K}$, calculate $\boldsymbol{\Phi}$ and $\boldsymbol{\Xi}$;
7: Calculate three monitoring statistics by (23), (25), (26);
8: Calculate thresholds by KDE, labeled as J_{th,T^2}, $J_{th,SPE}$ and J_{th,S^2}.

Algorithm 2 Online monitoring procedure of PSFA–EWC

1: Collect the test data $\boldsymbol{x}$, preprocess $\boldsymbol{x}$ by $\boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}_K$;
2: Calculate the latent variable by (22) and prediction error by (24);
3: Calculate three monitoring statistics by (23), (25), (26);
4: Judge the operating condition:
   a) Normal, return to step 1;
   b) A new mode appears, let $K=K+1$, return to Algorithm 1 to update the monitoring model;
   c) A fault occurs and the alarm is triggered.

The off-line training and online monitoring procedures have been summarized in Algorithm 1 and Algorithm 2, respectively. Fault detection rates (FDRs) and false alarm rates (FARs) are adopted to evaluate the performance.

B. Comparative design

RSFA [11], PCA–EWC [32], IMPPCA [19] and MCVA [18] are selected as the comparative methods in Table I. For PSFA–EWC, PSFA and RSFA, three monitoring statistics are calculated, where S^2 statistic is beneficial to distinguishing real faults and normal dynamic behaviors in multimode nonstationary processes. The remaining methods calculate two statistics and cannot separate real faults from normal variations. Assume that data from each mode are collected sequentially, the performance is evaluated by monitoring the current and the previously learned modes.

PSFA–EWC, RSFA and PCA–EWC can be regarded as adaptive methods, which avoid storing data and alleviating storage requirement. For Situations 1–11, PSFA and PSFA–EWC are compared to illustrate the catastrophic forgetting issue of PSFA and the continual learning ability of PSFA–EWC for successive nonstationary modes. When a new mode is identified by S^2 statistic and limited prior knowledge, a small set of normal data are collected and the model is updated off-line by extracting new information while consolidating the learned knowledge. PSFA–EWC furnishes the backward and forward transfer ability, namely, the updated PSFA–EWC model is able to monitor the previous modes and the learned knowledge is valuable to learn future new relevant modes. Equivalently, the simulation results of Situations 2, 3, 6–8 should be excellent. Conversely, the results of Situations 5, 10 and 11 are expected to be poor. The RSFA model is updated in real time and desired to track the system adaptively, as Situations 12–14 illustrated. For Situations 15–20, the design

TABLE I
COMPARATIVE SCHEMES

	Methods	Training sources (Model + Data)	Model label	Testing sources
Situation 1	PSFA	$\mathcal{M}_1$	A	$\mathcal{M}_1$
Situation 2	PSFA–EWC	A + $\mathcal{M}_2$	B	$\mathcal{M}_2$
Situation 3	PSFA–EWC	-	B	$\mathcal{M}_1$
Situation 4	PSFA	$\mathcal{M}_2$	C	$\mathcal{M}_2$
Situation 5	PSFA	-	C	$\mathcal{M}_1$
Situation 6	PSFA–EWC	B + $\mathcal{M}_3$	D	$\mathcal{M}_3$
Situation 7	PSFA–EWC	-	D	$\mathcal{M}_1$
Situation 8	PSFA–EWC	-	D	$\mathcal{M}_2$
Situation 9	PSFA	$\mathcal{M}_3$	E	$\mathcal{M}_3$
Situation 10	PSFA	-	E	$\mathcal{M}_1$
Situation 11	PSFA	-	E	$\mathcal{M}_2$
Situation 12	RSFA	$\mathcal{M}_1$	F	$\mathcal{M}_1$
Situation 13	RSFA	F + $\mathcal{M}_2$	G	$\mathcal{M}_2$
Situation 14	RSFA	G + $\mathcal{M}_3$	H	$\mathcal{M}_3$
Situation 15	PCA	$\mathcal{M}_1$	I	$\mathcal{M}_1$
Situation 16	PCA–EWC	I + $\mathcal{M}_2$	J	$\mathcal{M}_2$
Situation 17	PCA–EWC	-	J	$\mathcal{M}_1$
Situation 18	PCA–EWC	J + $\mathcal{M}_3$	L	$\mathcal{M}_3$
Situation 19	PCA–EWC	-	L	$\mathcal{M}_1$
Situation 20	PCA–EWC	-	L	$\mathcal{M}_2$
Situation 21	IMPPCA	$\mathcal{M}_1$, $\mathcal{M}_2$	M	$\mathcal{M}_1$
Situation 22	IMPPCA	-	M	$\mathcal{M}_2$
Situation 23	IMPPCA	$\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$	N	$\mathcal{M}_1$
Situation 24	IMPPCA	-	N	$\mathcal{M}_2$
Situation 25	IMPPCA	-	N	$\mathcal{M}_3$
Situation 26	MCVA	$\mathcal{M}_1$, $\mathcal{M}_2$	O	$\mathcal{M}_1$
Situation 27	MCVA	-	O	$\mathcal{M}_2$
Situation 28	MCVA	$\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$	P	$\mathcal{M}_1$
Situation 29	MCVA	-	P	$\mathcal{M}_2$
Situation 30	MCVA	-	P	$\mathcal{M}_3$

process of PCA–EWC is similar to that of PSFA–EWC. PCA–EWC is desired to provide the continual learning ability comparable to PSFA–EWC.

IMPPCA and MCVA are multiple-model methods, where the mode is identified and local models are built within each mode. Data from all possible modes are required before learning. When a novel mode arrives, sufficient samples should be collected and the model needs to be retrained on the entire dataset. IMPPCA and MCVA should provide excellent performance for Situations 21–30. However, it is intractable and time-consuming to collect complete data in practical systems [14]. The computational resources would increase for each retraining with the increasing number of modes.

V. CASE STUDIES

A. CSTH case

The CSTH process is a nonlinear nonstationary process and widely utilized as a benchmark for multimode process monitoring [14], [22]. Thornhill *et al.* built the CSTH model and the detail information was described in [37]. CSTH aims to mix the hot and cold water with desirable settings. Level, temperature and flow are manipulated by PI controllers. Six critical variables are selected for monitoring and three successive modes are considered in Table II. The numbers of training and testing samples are denoted as NoTrS and NoTeS, respectively. A random fault occurs in the level from the $501th$ sample and the fault amplitude is 0.15.

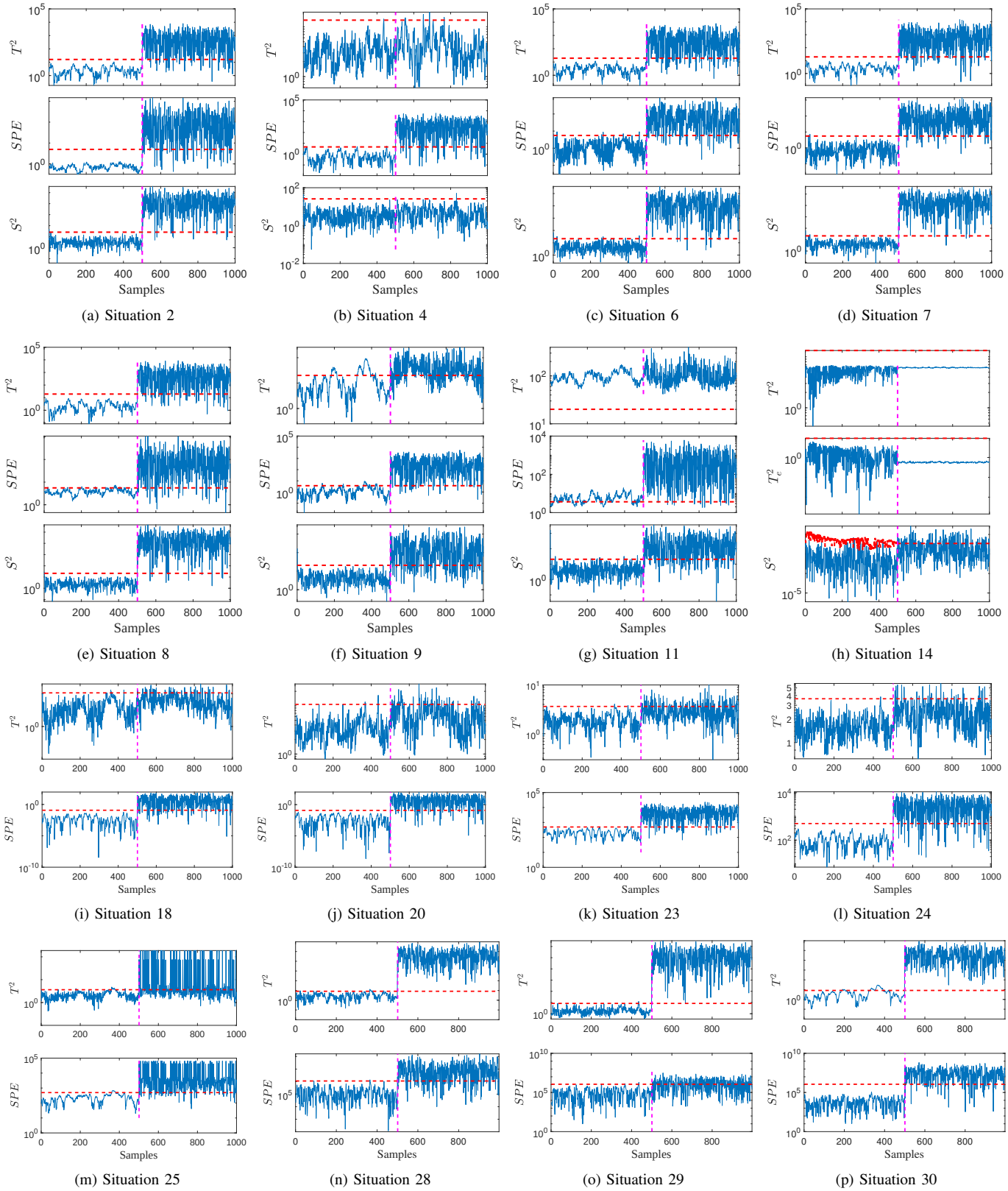
The monitoring results are summarized in Tables III and IV. Partial monitoring charts are depicted in Fig. 2 owing to paper length. Generally, PSFA–EWC provides excellent performance for sequential modes, where the real fault and normal variations can be distinguished by S^2 statistic. When a new mode is identified, 300 normal samples are collected and the PSFA–EWC model is updated based on these limited data, which could provide excellent performance for the current mode. For instance, the performance of Situations 2 and 6 is excellent and the FDRs of S^2 statistic are not less than 98%, which indicates that the fault is detected accurately by PSFA–EWC and reflects the forward transfer learning ability. Meanwhile, the previously learned knowledge is still consolidated while extracting new features, which is sufficient to monitor the past modes. Specifically, with regard to S^2 statistic, the FDRs of Situations 3, 7 and 8 are higher than 98%, which can illustrate the backward transfer learning ability of the proposed method. The FDRs of T^2 or SPE are similar for Situations 1–11. However, the FARs of Situation 11 are higher than 70%, which indicates that the significant knowledge of previous mode $\mathcal{M}_2$ is forgotten catastrophically. Succinctly, PSFA–EWC can transfer knowledge between modes, while it is difficult to establish an accurate PSFA model based on limited data.

For Situations 12–14, RSFA fails to monitor successive modes based on an adaptive model, where the FDRs of S^2 are less than 60%. RSFA is difficult to track the dramatic variations on the entire dataset. Analogous to PSFA–EWC, PCA–EWC is expected to offer prominent performance for sequential modes. However, the FDRs of Situations 16–20 cannot compare to the corresponding situations of PSFA–EWC. Although both methods utilize EWC to preserve the previously learned knowledge, PSFA can deal with dynamic slow features and S^2 is designed to reflect the unusual dynamic behaviors, while PCA is suitable to stationary data in each mode. IMPPCA and MCVA build the local models in each mode and the model needs to be retrained on the entire dataset when a new mode arrives. They deliver outstanding monitoring consequences for the learned modes, expect for Situation 30.

Generally, PSFA–EWC outperforms others for sequential modes, where the number of modes and samples per mode are not required in advance. When a new mode is identified, a few data are collected and the model is rapidly updated by assimilating new information while consolidating the learned knowledge. The RSFA model is updated when a new normal sample arrives, but fails to distinguish the normal changes and real faults in multimode nonstationary processes. Compared with PSFA–EWC, PCA–EWC is effective to detect the abnor-

TABLE II
NORMAL OPERATING MODES AND DATA INFORMATION OF CSTH

	Normal operating setting			Data information	
Mode number	Level SP	Temperature SP	Hot water valve	NoTrS	NoTeS
$\mathcal{M}_1$	9	10.5	4.5	1000	1000
$\mathcal{M}_2$	12	8	4	300	1000
$\mathcal{M}_3$	12	10.5	5.5	300	1000



(a) Situation 2 (b) Situation 4 (c) Situation 6 (d) Situation 7

(e) Situation 8 (f) Situation 9 (g) Situation 11 (h) Situation 14

(i) Situation 18 (j) Situation 20 (k) Situation 23 (l) Situation 24

(m) Situation 25 (n) Situation 28 (o) Situation 29 (p) Situation 30

Fig. 2. Monitoring charts of the CSTH case

TABLE III
FDRs (%) AND FARs (%) FOR PSFA, PSFA–EWC AND RSFA

	Method	CSTH						Pulverizing system					
		T^2		SPE		S^2		T^2		SPE		S^2	
		FDR	FAR	FDR	FAR	FDR	FAR	FDR	FAR	FDR	FAR	FDR	FAR
Situation 1	PSFA	43.6	7.4	95.0	1.4	77.0	0.2	99.92	2.15	99.92	0	20.05	0.86
Situation 2	PSFA–EWC	91.2	0	89.6	0	98.0	0.2	100	9.74	100	4.07	94.71	3.59
Situation 3	PSFA–EWC	92.0	0	92.4	0	99.2	9.8	99.92	1.29	99.92	1.29	94.01	14.04
Situation 4	PSFA	1.2	0	93.4	0.6	0.2	0.2	100	48.80	100	5.91	93.25	0.40
Situation 5	PSFA	4.2	6.2	95.0	0.8	20.0	21.8	100	15.19	99.92	0	93.13	14.04
Situation 6	PSFA–EWC	88.6	0	89.4	1.2	98.4	1.6	100	3.65	100	0.20	94.73	3.55
Situation 7	PSFA–EWC	90.6	0	93.4	0.8	99.0	0.8	99.92	1.00	99.92	0	88.10	10.89
Situation 8	PSFA–EWC	89.8		92.0	5.8	98.2	0.2	100	7.91	99.45	0.24	95.26	0.88
Situation 9	PSFA	68.8	23.2	94.4	8.0	78.8	0.2	100	1.32	100	0.20	53.63	2.13
Situation 10	PSFA	54.0	3.6	95.4	5.6	83.4	0.4	99.92	66.05	99.92	0.29	83.95	9.60
Situation 11	PSFA	100	100	96.4	74.4	83.2	2.4	100	79.63	99.45	0	91.79	0.24
Situation 12	RSFA	0	0	0	0	42.0	0.6	60.30	0	0	0	43.13	1.86
Situation 13	RSFA	0	0	0	0	59.2	1.0	84.49	0	0	0	37.59	1.68
Situation 14	RSFA	0	0	0	0	28.8	0.2	0	0	1.98	0	6.59	1.02

mality from static features but difficult to identify a new mode. PSFA–EWC, RSFA and PCA–EWC have the basically fixed model capacity, where a single model is updated continually. For IMPPCA and MCVA, the model is rebuilt based on the entire dataset when a new mode arrives and the model complexity would increase with the emergence of novel modes.

B. The pulverizing system

We focus on the coal pulverizing system of the 1030–MW ultra-supercritical thermal power plant in China [32]. The structure is depicted in Fig. 3, which is composed of coal feeder, coal mill, rotary separator, raw coal hopper and stone coal scuttle. The coal pulverizing system grinds the raw coal into pulverized coal with desired fineness and optimal temperature. According to the historical recording, the fault in outlet temperature occurs frequently and it is essential to investigate this sort of fault. Data from three successive modes are selected to illustrate the effectiveness, as listed in Table V. When a new mode arrives, only 540 normal samples are collected to update the model. The variables are selected by expert experience and prior knowledge.

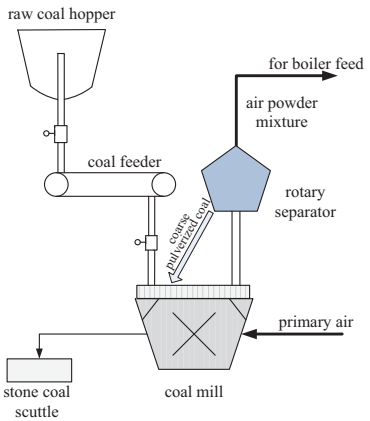


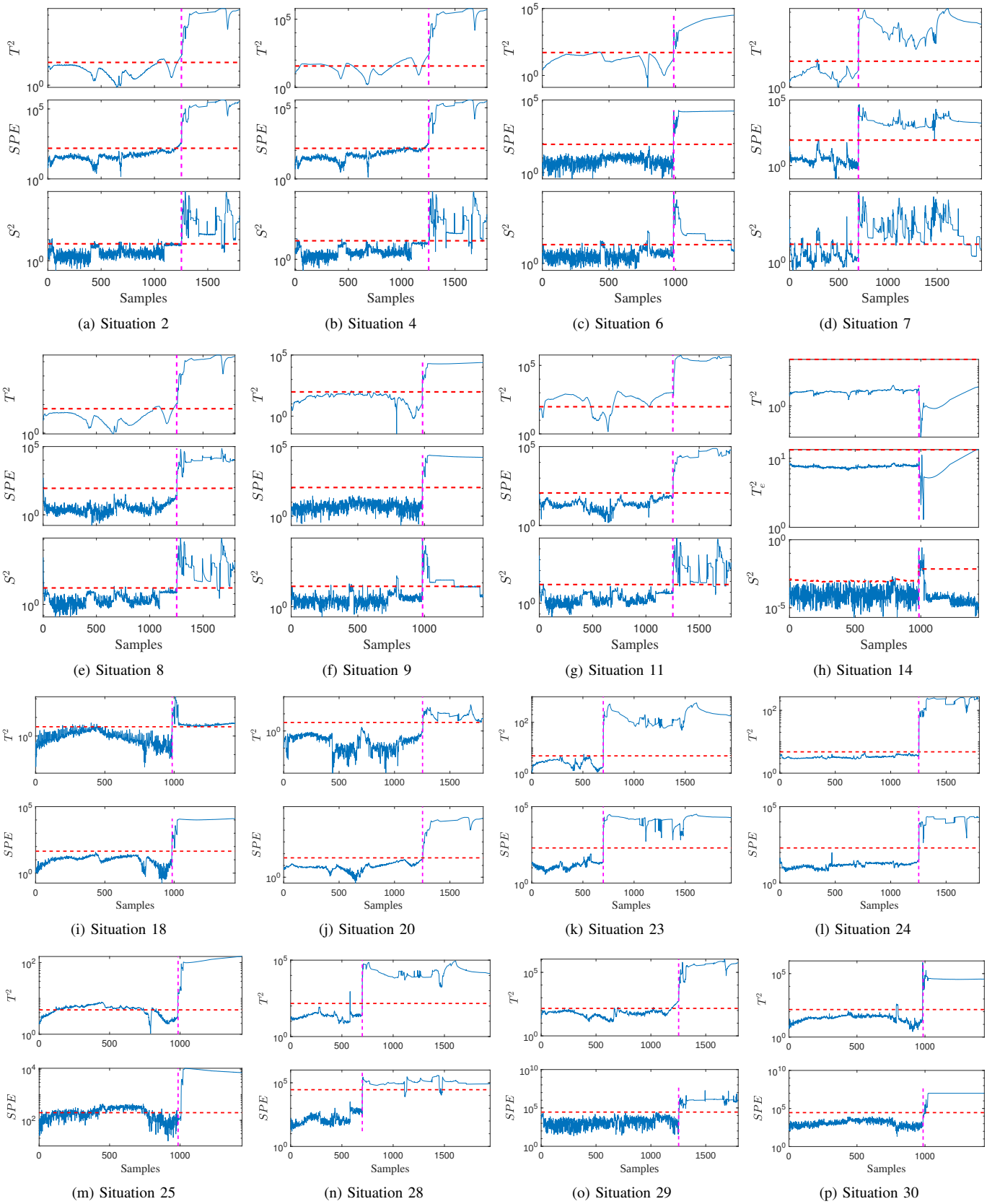
Fig. 3. Schematic diagram of coal pulverizing system

TABLE V
EXPERIMENTAL DATA OF THE PRACTICAL COAL PULVERIZING SYSTEM

Mode number	NoTrS	NoTeS	Fault location	Fault cause
$\mathcal{M}_1$	2520	1440	699	Pulverizer deflagration
$\mathcal{M}_2$	540	1800	1253	Hot primary air electric damper failure
$\mathcal{M}_3$	540	1440	986	Air leakage at primary air interface

The simulation results of 30 situations are summarized in Table III and Table IV. Partial monitoring charts are described in Fig. 4. With regard to S^2 statistic, the FDRs of Situations 6 and 9 are 94.73% and 53.63%, which indicates that the perviously learned knowledge from modes $\mathcal{M}_1$ and $\mathcal{M}_2$ is conducive to monitor mode $\mathcal{M}_3$. This phenomenon can reflect the forward transfer learning ability of PSFA–EWC. For T^2 statistic, the FARs of Situations 4, 10 and 11 are higher than 48%, while the FARs of PSFA–EWC are lower than 10%. PSFA suffers from catastrophic forgetting issue, where the model for one mode may not provide excellent performance for another mode. RSFA cannot monitor the multiple modes accurately and the FDRs of S^2 are lower than 44%. Only the FDR of T^2 is 84.49% for Situation 13. PCA–EWC can detect the faults in successive modes timely and the FDRs approach 100%. IMPPCA and MCVA offer favorable monitoring performance and the FDRs are convincing, except for Situation 25.

In conclusion, PSFA–EWC is capable of monitoring sequential modes accurately and the fault is confirmed by S^2 statistic. The model is updated continually by extracting new information while preserving the learned knowledge, thus avoiding performance degradation for similar modes as before. RSFA is effective to deal with slowly time-varying data and thus fails to track the dramatic changes on the entire dataset. Similar to PSFA–EWC, PCA–EWC enables to monitor the multiple modes based on an updated model for this case. IMPPCA and MCVA are able to monitor the learned modes. In terms of detection accuracy, the model complexity and applications, PSFA–EWC is the most desirable among five typical methods.



(a) Situation 2 (b) Situation 4 (c) Situation 6 (d) Situation 7

(e) Situation 8 (f) Situation 9 (g) Situation 11 (h) Situation 14

(i) Situation 18 (j) Situation 20 (k) Situation 23 (l) Situation 24

(m) Situation 25 (n) Situation 28 (o) Situation 29 (p) Situation 30

Fig. 4. Monitoring charts of the pulverizing system

TABLE IV
FDRs (%) AND FARs (%) FOR PCA–EWC, IMPPCA AND MCVA

	Methods	CSTH				Pulverizing system			
		T^2		SPE		T^2		SPE	
		FDR	FAR	FDR	FAR	FDR	FAR	FDR	FAR
Situation 15	PCA	83.0	4.4	94.2	0.8	99.92	1.00	99.92	1.00
Situation 16	PCA–EWC	9.2	0.6	90.8	0	100	0	100	0.16
Situation 17	PCA–EWC	8.2	0.8	91.6	0	99.92	0.14	99.92	0
Situation 18	PCA–EWC	13.0	0.8	90.6	0	100	2.44	100	0.20
Situation 19	PCA–EWC	12.2	1.4	92.4	0	99.92	0.14	99.92	0
Situation 20	PCA–EWC	12.6	1.2	91.8	0	99.45	0	99.45	0
Situation 21	IMPPCA	24.4	0.8	94.8	0.4	99.92	0	99.84	0.29
Situation 22	IMPPCA	0.8	0.2	91.0	0	99.45	0.48	99.45	4.23
Situation 23	IMPPCA	20.8	0.8	94.8	0.4	99.92	0.29	99.92	0
Situation 24	IMPPCA	8.4	0.2	89.0	0	99.45	0	99.45	0
Situation 25	IMPPCA	43.0	5.2	91.0	4.2	100	63.96	95.16	47.72
Situation 26	MCVA	100	3.0	77.71	0.2	100	0.14	0	0
Situation 27	MCVA	99.8	1.6	39.56	0.2	49.27	0	97.50	0
Situation 28	MCVA	100	3.0	80.92	0.6	100	0.14	96.07	0
Situation 29	MCVA	99.8	1.2	49.20	1.0	100	6.07	99.63	0
Situation 30	MCVA	100	16.8	87.75	0.2	100	1.62	97.34	0

VI. CONCLUSION

This paper has introduced a multimode PSFA algorithm with continual learning ability for multimode nonstationary process monitoring. The proposed PSFA–EWC method has powerful probabilistic interpretability and ability to deal with the measurement noise. When a new mode arrives, assume that a small set of data are collected, the single model is updated by consolidating new information while preserving the learned features. The previously learned knowledge is retained and may be beneficial to establishing an accurate model for future relevant modes, thus delivering backward and forward transfer learning ability. The PSFA features are extracted to form meaningful statistics for fault detection covering multimodes with only using recent mode data, with low storage and computational costs. Compared with several state-of-the-art methods, the effectiveness of PSFA–EWC is illustrated by a CSTH case and a practical coal pulverizing system.

The proposed PSFA–EWC algorithm requires that data from multiple modes have a certain degree of similarity, where the previously learned knowledge may be efficient for future modes. In future work, we would investigate the multimode nonstationary modes with applications to chemical systems, industrial manufacturing systems, etc. Besides, replay continual learning would be investigated for multimode process monitoring, where the modes are allowed to be diverse and the long-term continual learning ability is desired.

APPENDIX

A. Estimation of Fisher information matrix with PSFA

In order to approximate the posterior probability $P\left(\theta|\mathcal{M}_{i=1}^{K-1}\right)$, sequentially with incoming modes $K=2,\ldots,$ Laplace approximation [38], [39] is employed, i.e., local Gaussian probability density is used for its approximation centered at maximum posterior probability $\theta^*_{\mathcal{M}_{K-1}}$, with covariance of the gradient of the model's log likelihood function with respect to $\theta^*_{\mathcal{M}_{K-1}}$. The Fisher information matrix is the covariance of the gradient of the model's log likelihood function with respect to the local optimum, namely,

$$\begin{aligned}\boldsymbol{F} &= \mathbb{E}_{P_{\boldsymbol{x},\boldsymbol{y}}}\left[\nabla \log\ P\left(\boldsymbol{x},\boldsymbol{y}|\theta\right)\nabla \log\ P(\boldsymbol{x},\boldsymbol{y}|\theta)^T\right] \\ &= \frac{1}{T}\sum_t\left[\nabla \log\ P\left(\boldsymbol{x}_t,\boldsymbol{y}_t|\theta\right)\nabla \log\ P(\boldsymbol{x}_t,\boldsymbol{y}_t|\theta)^T\right]\end{aligned} \tag{27}$$

where $\theta=\theta^*_{\mathcal{M}_{K-1}}$ after the mode $\mathcal{M}_{K-1}$ has been learned. The conditional probability is calculated by

$$P\left(\boldsymbol{x}_t,\boldsymbol{y}_t|\theta\right)=P\left(\boldsymbol{x}_t|\boldsymbol{y}_t,\theta_x\right)P\left(\boldsymbol{y}_t|\theta_y\right)$$

Within the context of our PSFA model (1), parameters $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ are considered to calculate the corresponding Fisher information matrices, since the Laplacian is based on well-behaved function approximation which may not be applicable to noise. Besides, it is reasonable to assume that noise from multiple modes is independent and variance of unknown noise is constant in our problem. The gradient with regard to $\boldsymbol{V}$ is

$$\begin{aligned}\nabla_{\boldsymbol{V}} \log\ P\left(\boldsymbol{x}_t,\boldsymbol{y}_t|\theta\right) &= \frac{\partial \log\ P\left(\boldsymbol{x}_t|\boldsymbol{y}_t,\theta_x\right)}{\partial \boldsymbol{V}} \\ &= \boldsymbol{\Sigma}_x^{-1}\left(\boldsymbol{V}\boldsymbol{y}_t-\boldsymbol{x}_t\right)\boldsymbol{y}_t^T\end{aligned} \tag{28}$$

When the mode $\mathcal{M}_K$ has been learned, the Fisher information matrix about $\boldsymbol{V}$ is computed by

$$\begin{aligned}&\boldsymbol{F}^V_{\mathcal{M}_K} \\ =&\frac{1}{T_K}\sum_t \boldsymbol{\Sigma}_x^{-1}\left(\boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{y}_t-\boldsymbol{x}_t\right)\boldsymbol{y}_t^T\boldsymbol{y}_t\left(\boldsymbol{V}_{\mathcal{M}_K}\boldsymbol{y}_t-\boldsymbol{x}_t\right)^T\boldsymbol{\Sigma}_x^{-1}\end{aligned} \tag{29}$$

Since $\boldsymbol{\Sigma} = \boldsymbol{I} - \boldsymbol{\Lambda}^2$, $\boldsymbol{y}_t \sim N\left(\boldsymbol{\Lambda}\boldsymbol{y}_{t-1}, \boldsymbol{\Sigma}\right)$, and $\boldsymbol{\Lambda} = diag\left(\lambda_1, \ldots, \lambda_p\right)$, the gradient with respect to λ_i is

$$\begin{aligned}
&\nabla_{\lambda_i} \log\ P\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\theta\right) \\
=&\frac{\partial \log\ P\left(\boldsymbol{y}_t|\theta_y\right)}{\partial \lambda_i} \\
=&\frac{-\lambda_i^3 + y_{t,i}y_{t-1,i}\lambda_i^2 + \left(1 - y_{t,i}^2 - y_{t-1,i}^2\right)\lambda_i + y_{t,i}y_{t-1,i}}{\left(1-\lambda_i^2\right)^2} \\
\triangleq& g\left(y_{t,i}, y_{t-1,i}, \lambda_i\right)
\end{aligned} \tag{30}$$

For mode $\mathcal{M}_K$, the Fisher information matrix about λ_i is

$$F_{\lambda_i} = \frac{1}{T_K}\sum_t g\left(y_{t,i}, y_{t-1,i}, \lambda_{\mathcal{M}_K,i}\right)^2, i = 1, \cdots, p \tag{31}$$

where $\lambda_{\mathcal{M}_K,i}$ is the ith element of diagonal matrix $\boldsymbol{\Lambda}_{\mathcal{M}_K}$, $\boldsymbol{F}_{\mathcal{M}_K}^{\Lambda} = diag\left(F_{\lambda_1}, \cdots, F_{\lambda_p}\right)$.

B. Estimation of sufficient statistics

Similar to [34], Kalman filter and Tanch-Tung-Striebel (RTS) smoother [40] are adopted, which contains the forward and backward recursion steps.

First, the forward recursions are adopted to estimate the posterior distribution $P\left(\boldsymbol{y}_t|\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \theta^{old}\right) \sim N\left(\boldsymbol{\mu}_t, \boldsymbol{U}_t\right)$ sequentially. The posterior marginal is calculated by

$$\begin{aligned}
&\int N\left(\boldsymbol{y}_{t-1}|\boldsymbol{\mu}_{t-1}, \boldsymbol{U}_{t-1}\right) N\left(\boldsymbol{y}_t|\boldsymbol{\Lambda}\boldsymbol{y}_{t-1}, \boldsymbol{\Sigma}\right) d\boldsymbol{y}_{t-1} \\
&= N\left(\boldsymbol{y}_t|\boldsymbol{\Lambda}\boldsymbol{y}_{t-1}, \boldsymbol{P}_{t-1}\right)
\end{aligned}$$

where $\boldsymbol{P}_{t-1}$ is the variance.

Then, parameters of the posterior distribution $P\left(\boldsymbol{Y}_K|\boldsymbol{X}_K, \theta^{old}\right)$ are acquired by backward recursion steps. The procedure is summarized in Algorithm 3.

Algorithm 3 E-step in PSFA-EWC

Input: $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Lambda}$, $\boldsymbol{V}$, $\boldsymbol{X}_K$
Output: $\mathbb{E}\left[\boldsymbol{y}_t|\boldsymbol{X}_K\right]$, $\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_{t-1}^T|\boldsymbol{X}_K\right]$, $\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^T|\boldsymbol{X}_K\right]$

% Forward steps by Kalman filter:
Initialize $\boldsymbol{K}_1 = \boldsymbol{\Sigma}_1\boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{\Sigma}_1\boldsymbol{V}^T + \boldsymbol{\Sigma}_x\right)^{-1}$, $\boldsymbol{\mu}_1 = \boldsymbol{K}_1\boldsymbol{x}_1$, $\boldsymbol{U}_1 = \left(\boldsymbol{I} - \boldsymbol{K}_1\boldsymbol{V}\right)\boldsymbol{\Sigma}_1$
for $t = 1 : T_K$ **do**
  $\boldsymbol{P}_{t-1} = \boldsymbol{\Lambda}\left(\boldsymbol{U}_{t-1} - \boldsymbol{I}\right)\boldsymbol{\Lambda}^T + \boldsymbol{I}$
  $\boldsymbol{K}_t = \boldsymbol{P}_{t-1}\boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{P}_{t-1}\boldsymbol{V}^T + \boldsymbol{\Sigma}_x\right)^{-1}$
  $\boldsymbol{\mu}_t = \boldsymbol{\Lambda}\boldsymbol{\mu}_{t-1} + \boldsymbol{K}_t\left(\boldsymbol{x}_t - \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{\mu}_{t-1}\right)$
  $\boldsymbol{U}_t = \left(\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{V}\right)\boldsymbol{P}_{t-1}$
end for
% Backward steps by RTS smoother
Initialize $\hat{\boldsymbol{\mu}}_{T_K} = \boldsymbol{\mu}_{T_K}$, $\hat{\boldsymbol{U}}_{T_K} = \boldsymbol{U}_{T_K}$
for $t = T_K : 2$ **do**
  $\boldsymbol{J}_{t-1} = \boldsymbol{U}_{t-1}\boldsymbol{\Lambda}^T\boldsymbol{P}_{t-1}^{-1}$
  $\hat{\boldsymbol{\mu}}_{t-1} = \boldsymbol{\mu}_{t-1} + \boldsymbol{J}_{t-1}\left(\hat{\boldsymbol{\mu}}_t - \boldsymbol{\Lambda}\boldsymbol{\mu}_{t-1}\right)$
  $\hat{\boldsymbol{U}}_{t-1} = \boldsymbol{U}_{t-1} + \boldsymbol{J}_{t-1}\left(\hat{\boldsymbol{U}}_t - \boldsymbol{P}_{t-1}\right)\boldsymbol{J}_{t-1}^T$
end for
% Calculate the sufficient statistics
for $t = 1 : T_K$ **do**
  $\mathbb{E}\left[\boldsymbol{y}_t|\boldsymbol{X}_K\right] = \hat{\boldsymbol{\mu}}_t$
  $\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_{t-1}^T|\boldsymbol{X}_K\right] = \boldsymbol{J}_{t-1}\hat{\boldsymbol{U}}_t + \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_{t-1}^T$
  $\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t^T|\boldsymbol{X}_K\right] = \hat{\boldsymbol{U}}_t + \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_t^T$
end for

REFERENCES

[1] T. J. Rato, J. Blue, J. Pinaton, and M. S. Reis, "Translation-invariant multiscale energy-based PCA for monitoring batch processes in semiconductor manufacturing," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 894–904, Apr. 2017.

[2] J. Shi, J. Sun, Y. Yang, and D. Zhou, "Distributed self-triggered formation control for multi-agent systems," *Sci. China Ser. F*, vol. 63, no. 10, pp. 1–3, 2020.

[3] Y. Qin, Y. Yan, H. Ji, and Y. Wang, "Recursive correlative statistical analysis method with sliding windows for incipient fault detection," *IEEE Trans. Ind. Electron.*, vol. 69, no. 4, pp. 4185–4194, Apr. 2022.

[4] R. Sun and Y. Wang, "C-IPLS-IKPLS for modeling and detecting nonlinear multimode processes," *Ind. Eng. Chem. Res.*, vol. 60, no. 4, pp. 1684–1698, 2021.

[5] S. J. Qin, Y. Dong, Q. Zhu, J. Wang, and Q. Liu, "Bridging systems theory and data science: A unifying review of dynamic latent variable analytics and process monitoring," *Annu. Rev. Control*, vol. 50, pp. 29–48, 2020.

[6] Y. Jiang and S. Yin, "Recursive total principle component regression based fault detection and its application to vehicular cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1415–1423, Apr. 2018.

[7] Y. Hu and C. Zhao, "Fault diagnosis with dual cointegration analysis of common and specific nonstationary fault variations," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 237–247, Jan. 2020.

[8] X. Ma, Y. Si, Y. Qin, and Y. Wang, "Fault detection for dynamic processes based on recursive innovational component statistical analysis," *IEEE Trans. Autom. Sci. Eng.*, to be published, doi: 10.1109/TASE.2022.3149591.

[9] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, 2002.

[10] C. Shang, F. Yang, X. Gao, X. Huang, J. A. Suykens, and D. Huang, "Concurrent monitoring of operating condition deviations and process dynamics anomalies with slow feature analysis," *AIChE J.*, vol. 61, no. 11, pp. 3666–3682, 2015.

[11] C. Shang, F. Yang, B. Huang, and D. Huang, "Recursive slow feature analysis for adaptive monitoring of industrial processes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 11, pp. 8895–8905, Nov. 2018.

[12] W. Yu and C. Zhao, "Recursive exponential slow feature analysis for fine-scale adaptive processes monitoring with comprehensive operation status identification," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3311–3323, Jun. 2019.

[13] F. Guo, C. Shang, B. Huang, K. Wang, F. Yang, and D. Huang, "Monitoring of operating point and process dynamics via probabilistic slow feature analysis," *Chemometr. Intell. Lab. Syst.*, vol. 151, no. 151, pp. 115–125, 2016.

[14] M. Quiñones-Grueiro, A. Prieto-Moreno, C. Verde, and O. Llanes-Santiago, "Data-driven monitoring of multimode continuous processes: A review," *Chemometr. Intell. Lab. Syst.*, vol. 189, pp. 56–71, 2019.

[15] H. Ma, Y. Hu, and H. Shi, "A novel local neighborhood standardization strategy and its application in fault detection of multimode processes," *Chemometr. Intell. Lab. Syst.*, vol. 118, pp. 287–300, 2012.

[16] J. Shang and M. Chen, "Recursive dynamic transformed component statistical analysis for fault detection in dynamic processes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 1, pp. 578–588, Jan. 2018.

[17] W. Shao, Z. Ge, L. Yao, and Z. Song, "Bayesian nonlinear Gaussian mixture regression and its application to virtual sensing for multimode industrial processes," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 871–885, Apr. 2020.

[18] Q. Wen, Z. Ge, and Z. Song, "Multimode dynamic process monitoring based on mixture canonical variate analysis model," *Ind. Eng. Chem. Res.*, vol. 54, no. 5, pp. 1605–1614, 2015.

[19] J. Zhang, H. Chen, S. Chen, and X. Hong, "An improved mixture of probabilistic PCA for nonlinear data-driven process monitoring," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 198–210, Jan. 2019.

[20] L. Zhou, J. Zheng, Z. Ge, Z. Song, and S. Shan, "Multimode process monitoring based on switching autoregressive dynamic latent variable model," *IEEE Trans. Ind. Electron.*, vol. 65, no. 10, pp. 8184–8194, Oct. 2018.

[21] R. Tan, J. R. Ottewill, and N. F. Thornhill, "Nonstationary discrete convolution kernel for multimodal process monitoring," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3670–3681, Sep. 2020.

[22] K. Huang, Y. Wu, C. Yang, G. Peng, and W. Shen, "Structure dictionary learning-based multimode process monitoring and its application to

aluminum electrolysis process," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1989–2003, Oct. 2020.
[23] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, and A. Grabska-Barwinska, "Overcoming catastrophic forgetting in neural networks," *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.
[24] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nat. Commun.*, vol. 11, no. 1, pp. 4069–4069, 2020.
[25] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends Cogn. Sci.*, vol. 24, no. 12, pp. 1028–1040, 2020.
[26] N. Y. Masse, G. D. Grant, and D. J. Freedman, "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization," *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 44, pp. E10467–E10475, 2018.
[27] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, 2019.
[28] J. Zhang, D. Zhou, M. Chen, and X. Hong, "Continual learning for multimode dynamic process monitoring with applications to an ultra-supercritical thermal power plant," *IEEE Trans. Autom. Sci. Eng.*, to be published, doi: 10.1109/TASE.2022.3144288.
[29] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE T. Pattern Anal.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.
[30] Z. Chen and B. Liu, "Lifelong machine learning," *Synthesis Lectures on Artificial Intelligence & Machine Learning*, vol. 12, no. 3, pp. 1–207, 2018.
[31] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, 2016.
[32] J. Zhang, D. Zhou, and M. Chen, "Monitoring multimode processes: a modified PCA algorithm with continual learning ability," *J. Process Contr.*, vol. 103, pp. 76–86, 2021.
[33] R. Turner and M. Sahani, "A maximum-likelihood interpretation for slow feature analysis," *Neural Comput*, vol. 19, no. 4, pp. 1022–1038, 2007.
[34] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, and M. Pantic, "Probabilistic slow features for behavior analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 5, pp. 1034–1048, May 2016.
[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B.*, vol. 39, no. 1, pp. 1–22, 1977.
[36] F. Huszár, "On quadratic penalties in elastic weight consolidation," *arXiv preprint arXiv:1712.03847*, 2017.
[37] N. F. Thornhill, S. C. Patwardhan, and S. L. Shah, "A continuous stirred tank heater simulation model with applications," *J. Process Contr.*, vol. 18, no. 3, pp. 347–360, 2008.
[38] R. Aljundi, *Continual Learning in Neural Networks*. PhD thesis, KU Leuven, 2019.
[39] J. Martens, "New insights and perspectives on the natural gradient method," *J. Mach. Learn. Res.*, vol. 21, no. 146, pp. 1–76, 2020.
[40] S. Sarkka, "Unscented rauch–tung–striebel smoother," *IEEE Trans. Automat. Contr.*, vol. 53, no. 3, pp. 845–849, Apr. 2008.

Jingxin Zhang received B.E. degree in Electrical Engineering and Automation from Harbin Engineering University, Harbin, China, the M.E. degree in Control Science and Engineering from Harbin Institute of Technology, Harbin, China, in 2014 and 2016, respectively, and the Ph.D. degree in Control Science and Engineering from Tsinghua University, Beijing, China, in 2022. She is currently a Lecturer with the School of Automation, Southeast University, Nanjing, China. Her research interests include continual learning, data-driven fault detection and diagnosis, multimode process monitoring, performance monitoring and their applications in industrial processes.

Donghua Zhou (SM'99-F'19, IEEE) received the B.Eng., M. Sci., and Ph.D. degrees all in electrical engineering from Shanghai Jiaotong University, China, in 1985, 1988, and 1990, respectively. He was an Alexander von Humboldt research fellow with the university of Duisburg, Germany from 1995 to 1996, and a visiting scholar with Yale university, USA from 2001 to 2002. He joined Tsinghua university in 1996, and was promoted as full professor in 1997, he was the head of the department of automation, Tsinghua university, during 2008 and 2015. He is now a vice president, Shandong University of Science and Technology, and a joint professor of Tsinghua university. He has authored and coauthored over 230 peer-reviewed international journal papers and 7 monographs in the areas of fault diagnosis, fault-tolerant control and operational safety evaluation. Dr. Zhou is a fellow of IEEE, CAA and IET, a member of IFAC TC on SAFEPROCESS, an associate editor of Journal of Process Control, the vice Chairman of Chinese Association of Automation (CAA) the TC Chair of the SAFEPROCESS committee, CAA. He was also the NOC Chair of the 6th IFAC Symposium on SAFEPROCESS 2006.

Maoyin Chen received the B.S. degree in mathematics and the M.S. degree in control theory and control engineering from Qufu Normal University, Shandong, China, in 1997 and 2000, respectively, and the Ph.D. degree in control theory and control engineering from Shanghai Jiaotong University, Shanghai, China, in 2003. From 2003 to 2005, he was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China. From 2006 to 2008, he visited Potsdam University, Potsdam, Germany, as an Alexander von Humboldt Research Fellow. Since October 2008, he has been an Associated Professor with the Department of Automation, Tsinghua University. He has authored and coauthored over 110 peer-reviewed international journal papers. He has won the first prize in natural science (2011, ranked first) and the second prize (2019, ranked first) of CAA. His research interests include fault prognosis and complex systems.

Xia Hong received the B.Sc. and M.Sc. degrees from the National University of Defense Technology, China, in 1984 and 1987, respectively, and the Ph.D. degree from The University of Sheffield, U.K., in 1998, all in automatic control. She was a Research Assistant with the Beijing Institute of Systems Engineering, Beijing, China, from 1987 to 1993. She was a Research Fellow with the Department of Electronics and Computer Science, University of Southampton, from 1997 to 2001.

She is currently a Professor with the Department of Computer Science, School of Mathematical, Physical and Computational Sciences, University of Reading. She is actively involved in research into nonlinear systems identification, data modeling, estimation and intelligent control, neural networks, pattern recognition, learning theory, and their applications. She has authored over 170 research papers, and co-authored a research book. Dr. Hong received the Donald Julius Groen Prize from IMechE in 1999.