Adaptive monitoring for multimode nonstationary processes using cointegration analysis and probabilistic slow feature analysis

Adaptive monitoring for multimode nonstationary processes using cointegration analysis and probabilistic slow feature analysis

Jingxin Zhang[a,b,*], Min Wang[c,*], Xu Xu[d], Donghua Zhou[e] and Xia Hong[f]

[a]School of Automation, Southeast University, Nanjing 210096, China
[b]MOE Key Laboratory of Measurement and Control of Complex Systems of Engineering, Nanjing 210096, China
[c]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[d]Yangze River Delta Information Intelligence Innovation Research Institute, Wuhu 314006, China
[e]Department of Automation, Tsinghua University, Beijing 100084, China
[f]Department of Computer Science, School of Mathematical, Physical and Computational Sciences, University of Reading, RG6 6AY, U.K.

ARTICLE INFO

Keywords:
Multimode nonstationary process monitoring
Recursive attention probabilistic slow feature analysis
Adaptive cointegration analysis
Continual learning

ABSTRACT

The condition monitoring of nonlinear, nonstationary and multimode processes is a difficult problem. Traditional multimode process monitoring methods generally assume that data from all potential modes are available, yet new modes may appear continuously in practice. This paper investigates an intelligent adaptive monitoring method for multimode nonstationary processes, which can deal with the appearance of new modes with ease. A full-condition comprehensive framework is proposed to decompose feature subspaces. First, long-term equilibrium features are extracted by adaptive cointegration analysis (ACA) to identify the mode, without using any prior mode information intelligently for online applications. Then, recursive attention probabilistic slow feature analysis integrated with elastic weight consolidation (RAttPSFA-EWC) is investigated to deal with the remaining dynamic information and extract dynamic and static slow features to maintain continual learning for multimodes. Once a new mode is detected automatically, the previously learned knowledge is consolidated while extracting new features, which is beneficial to enhancing the performance of similar modes. The proposed ACA-RAttPSFA-EWC acts as online adaptive method by parameter updates with incoming normal data. Furthermore, several advanced methods are compared to demonstrate the strengths of ACA-RAttPSFA-EWC, and the proposed method is validated to be effective using a numerical case and a practical system.

1. Introduction

In order to enhance the safety and reliability of industrial processes, process monitoring has been becoming essential and increasingly well researched [1, 2, 3, 4]. Owing to the switching operating points or raw materials, some industrial processes typically operate under multiple modes [5, 6, 7].

Multimode process monitoring methods can be classified as either multiple model methods [8, 9] or single model methods [10, 11, 12]. In a multiple model method, data are divided into several clusters and local monitoring models are built within each cluster. For instance, a common and specific feature extraction method was explored to monitor the multimode processes with common features [8]. In [13], common dictionary and mode-specific dictionaries were investigated for multiple modes and the mode was identified via the reconstruction error. Besides, a hierarchical Dirichlet process integrated with Hidden semi-Markov model was presented to settle the missing mode information issue [9]. Multiple model methods require complete data from all potential modes. In industrial applications, data are naturally nonstationary in each mode and/or novel modes may appear continuously, which implies that the monitoring model needs to be retrained, which is impractical.

Single model methods establish a single monitoring model for multiple modes, where the multimodal data are transformed into a unimodal distribution or the model parameters are updated recursively to adapt to the varying variations. Recursive slow feature analysis (RSFA) was presented to isolate temporal dynamics from steady conditions [12], which was beneficial to identifying modes. Subsequently, a recursive exponential slow feature analysis was developed to distinguish the normal slow changes and incipient faults [14]. Besides, an exponential analytic stationary subspace analysis was proposed for nonstationary process, which also could distinguish the real faults from normal changes while being robust to the disturbances [15]. Recursive cointegration analysis (RCA) was investigated for single-mode nonstationary processes [11] and the mode was identified automatically. However, in the case of nonstationary processes, traditional recursive methods may fail as these could not quickly track the dramatic variations between consecutive modes when they are applied to multiple modes.

Consider that data from multiple modes are collected sequentially, continual learning has been applied to multimode process monitoring [16, 17]. The concept of continual learning is to consolidate the previously learned knowledge [18, 19] while assimilating new features from new modes. The model can be learned continually with limited data and computational resources. According to the manner of preserving significant information, continual learning methods

*Corresponding author
jingxinzhang@seu.edu.cn (J. Zhang); mwang@uestc.edu.cn (M. Wang); xuxu@ustc.win (X. Xu); zdh@mail.tsinghua.edu.cn (D. Zhou); x.hong@reading.ac.uk (X. Hong)
ORCID(s):

are sorted into regularization-based [20], replay [21] and parameter isolation methods [22]. Recently an adaptive method was proposed using adaptive cointegration analysis (ACA), recursive principal component analysis (RPCA) and elastic weight consolidation (ACA-RPCA-EWC) [16], which could identify the mode automatically and track the rapid variations accurately in multimode nonstationary processes. However, it does not deal with measurement noise directly. To account for the uncertainty, probabilistic slow feature analysis with EWC (PSFA-EWC) was investigated for multimode nonstationary processes [17], where the measurement noise was considered and missing data were modeled with ease [23]. However, in the aforementioned regularization-based methods [10, 17], data from multiple modes are required to be similar in some sense with the mode information given as a priori.

Since almost all the aforementioned multimode process monitoring methods assume that data from all potential modes are available beforehand except for ACA-RPCA-EWC, it is highly desired to develop a mode-free method for multimode nonstationary processes. For replay-based continual learning, a few representative data are stored or pseudo-data are generated for each mode, which would be replayed when a new mode arrives. In [24], multimode nonlinear sparse dynamic inner principal component analysis was proposed to monitor diverse modes, where representative data from each mode were selected based on cosine similarity and would be integrated with the new data for retraining. Similar to [10, 17], the mode information should be available in advance.

In practical applications, such as the large-scale thermal power generation plants, it is intractable to obtain the accurate mode information. The process data are obviously nonstationary or stationary owing to the time-varying load and the coal type. Besides, the data distribution and the relationship between variables may also change because the components of the raw materials vary slowly with the environment. Meanwhile, since the plants generally operate under a high-pressure and high-speed rotating condition, the process data are easily affected by noise. Aforementioned methods cannot tackle this issue, where the relationship between variables changes, the mode information is unavailable and the noise should be considered simultaneously.

Against this background, this work introduces an adaptive method for multimode nonlinear nonstationary processes, which provides a comprehensive monitoring framework and could account for uncertainty due to probabilistic interpretation. The expert knowledge is only used to decompose the variables into several blocks and select the mode-sensitive variables for offline training. When a new mode is detected by ACA automatically, only an extremely small amount of data are collected for offline training and the previously learned knowledge is preserved to provide continual learning ability. Subsequently, the model parameters would be updated adaptively based on the forthcoming data, which is capable of tracking the dynamic variations accurately and can provide excellent monitoring performance for nonstationary processes. Moreover, different from most state-of-the-art monitoring methods [17, 8], the proposed method is free from storing historical data and incoming sequential data from different modes. Meanwhile, it can distinguish the real faults, nonstationarity and normal mode switching for online applications without much prior mode information.

The contributions of this paper are summarized below:

a) An adaptive monitoring framework is investigated for multimode nonstationary processes, where the mode is identified automatically without human intervention. To obtain optimal performance, variables are divided into three parts and feature subspaces are decomposed systematically to achieve a full-condition monitoring model. The long-term equilibrium features are extracted via ACA, which are used to identify the mode automatically.

b) Recursive attention PSFA with EWC (RAttPSFA-EWC) is investigated to track the slow variations adaptively, and the learned knowledge of previous modes is consolidated when a new mode is detected. RAttPSFA-EWC is introduced to deal with the remaining dynamic information that are unaccounted for by ACA, in which an attention mechanism is adopted to focus on the significant information and model nonlinearity. The measurement noise is considered using PSFA.

c) The proposed method is investigated based on the operating mechanism, expert knowledge and abundant data, which can provide satisfactory monitoring performance as well as excellent interpretability. The effectiveness of the proposed method is validated via a numerical case and a practical industrial case.

The rest of this paper is organized below. Section 2 explains the problem statement by reviewing the procedure of ACA, and introducing attention PSFA (AttPSFA) for a single nonlinear dynamic mode. Section 3 outlines the objective of AttPSFA-EWC and introduces the technical details of the proposed RAttPSFA-EWC for multimode processes. Then, the monitoring procedure is summarized and several advanced approaches are discussed. The effectiveness of ACA-RAttPSFA-EWC is demonstrated using a numerical case and a practical pulverizing system in Section 4. The conclusion is provided in Section 5.

2. Preliminaries

2.1. Problem statement

Assume that nonstationary data from multiple modes are received sequentially. To describe a single mode $\mathcal{M}_K$ ($K = 1, 2, \ldots$), let $\boldsymbol{X}_K \in \Re^{N_K \times m}$ be collected for offline training, where N_K is the number of samples and m is the number of variables. This work investigates an online adaptive monitoring method for multimode nonstationary processes based on cointegration analysis (CA) and AttPSFA, as outlined in Sections 2.2 and 2.4, respectively.

According to correlation analysis, prior knowledge and the augmented Dickey-Fuller (ADF) test [16], data $\boldsymbol{X}_K$ are decomposed into $\boldsymbol{X}_K \longrightarrow \left\{\boldsymbol{X}_{0,K}\ \boldsymbol{X}_{1,K}\ \boldsymbol{X}_{2,K}\right\}$. First, data $\boldsymbol{X}_{0,K}$ represent the stationary variables in each mode that are sensitive to mode switching, which are selected by the ADF test and prior knowledge. Then, data $\boldsymbol{X}_{1,K}$ denote the nonstationary ones that share similar trends, which are selected based on the process mechanism. The remaining nonstationary data $\boldsymbol{X}_{2,K}$ have no prominent role in any mode [16]. Note that the data decomposition may vary for multiple modes. Specifically, the dimension of data $\boldsymbol{X}_{0,K}$ is same for different modes, while the dimensions of $\boldsymbol{X}_{1,K}$ and $\boldsymbol{X}_{2,K}$ may be different.

For offline training procedure, CA extracts the long-term equilibrium features from $\boldsymbol{X}_{1,K}$, as described in Section 2.2. The rest information after CA together with $\boldsymbol{X}_{2,K}$ would be processed by AttPSFA-EWC in Section 3.1. For online applications, when the system operates normally, the model parameters are updated adaptively by ACA in Section 2.3 and RAttPSFA-EWC in Section 3.2. Note that data $\boldsymbol{X}_{0,K}$ are only used to calculate the monitoring statistics.

2.2. Cointegration analysis for mode $\mathcal{M}_K$

CA aims to deal with nonstationary time series data, which are stationary after being differentiated several times [11]. The linear combinations of aforementioned nonstationary variables (called cointegration variables) are stationary after the CA algorithm. Intuitively speaking, cointegration variables furnish the long-term equilibrium relationship, which would be broken when a real fault happens.

We refer to $\boldsymbol{X}_{1,K}$ as primary nonstationary signal. CA aims to extract long-term equilibrium features from nonstationary data $\boldsymbol{X}_{1,K}$ offline. Consider the primary nonstationary signal $\boldsymbol{X}_{1,K} \in \Re^{N_K \times m_1}$ at mode $\mathcal{M}_K$, consisting of N_K consecutive vector observations $\left\{\boldsymbol{x}_t\right\}_{t=1}^{N_K}$ with $\boldsymbol{x}_t \in \Re^{m_1}$ [25, 26]. To estimate the CA parameters, construct the vector error-correction (VEC) model below:

$$\Delta\boldsymbol{x}_t = \sum_{j=1}^{p_1-1} \boldsymbol{\Omega}_j \Delta\boldsymbol{x}_{t-j} + \boldsymbol{\Gamma}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t \tag{1}$$

in which $\Delta\boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$ and p_1 is the order of VEC model. $\boldsymbol{\epsilon}_t$ is the Gaussian white noise with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Xi})$. $\boldsymbol{\Gamma} = \boldsymbol{\Upsilon}\boldsymbol{W}_{f,K}^T \in \Re^{m_1 \times m_1}$, $\boldsymbol{\Upsilon} \in \Re^{m_1 \times r}$ and the cointegration matrix $\boldsymbol{W}_{f,K} \in \Re^{m_1 \times r}$ are of full rank r, and r is estimated by the trace test [26]. CA seeks to make the equilibrium errors, namely each column of $\boldsymbol{X}_{1,K}\boldsymbol{W}_{f,K}$, as stationary as possible.

To estimate $\boldsymbol{W}_{f,K}$, initially define the temporal difference vector $\Delta\boldsymbol{x}_{p_1+1} = \boldsymbol{x}_{p_1+1} - \boldsymbol{x}_{p_1}$ and the augmented vector $\Delta\boldsymbol{x}_1^{p_1} = \left[\Delta\boldsymbol{x}_1^T\ \Delta\boldsymbol{x}_2^T\ \cdots\ \Delta\boldsymbol{x}_{p_1}^T\right]^T \in \Re^{p_1 m_1}$. Then, construct the matrices $\Delta\boldsymbol{X}_{p_1} = \left[\Delta\boldsymbol{x}_{p_1+1}\ \Delta\boldsymbol{x}_{p_1+2}\ \cdots\ \Delta\boldsymbol{x}_{N_K}\right]^T \in \Re^{(N_K-p_1)\times m_1}$ and $\Delta\boldsymbol{X}^{p_1} = \left[\Delta\boldsymbol{x}_1^{p_1}\ \Delta\boldsymbol{x}_2^{p_1}\ \cdots\ \Delta\boldsymbol{x}_{N_K-p_1}^{p_1}\right]^T \in \Re^{(N_K-p_1)\times p_1 m_1}$. Two sets of the prediction errors $\tilde{\boldsymbol{E}}_0$ and $\tilde{\boldsymbol{E}}_1$ are defined according to

$$\tilde{\boldsymbol{E}}_0 = \Delta\boldsymbol{X}_{p_1} - \Delta\boldsymbol{X}^{p_1}\boldsymbol{\Theta} \tag{2}$$

$$\tilde{\boldsymbol{E}}_1 = \boldsymbol{X}_{p_1} - \Delta\boldsymbol{X}_1^p\boldsymbol{\Phi}. \tag{3}$$

Then, the regression parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are estimated using ordinary least squares, such that $\tilde{\boldsymbol{E}}_0^T\tilde{\boldsymbol{E}}_0$ and $\tilde{\boldsymbol{E}}_1^T\tilde{\boldsymbol{E}}_1$ are minimized for (2) and (3) respectively.

According to the Johansen test [26], $\boldsymbol{W}_{f,K}$ is estimated by solving the eigenvalue decomposition (EVD) problem

$$\left|\tilde{\lambda}\boldsymbol{S}_{11} - \boldsymbol{S}_{10}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}\right| = 0 \tag{4}$$

where $\boldsymbol{S}_{i,j} = \frac{1}{N-p_1}\tilde{\boldsymbol{E}}_i^T\tilde{\boldsymbol{E}}_j$, $i, j = 0, 1$, $\tilde{\lambda}$ is the corresponding eigenvalue of EVD problem. Subsequently, (4) is reformulated equivalently as

$$\boldsymbol{A}^{(K)}\boldsymbol{w} = \lambda\boldsymbol{B}^{(K)}\boldsymbol{w} \tag{5}$$

where $\boldsymbol{A}^{(K)} = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{S}_{01}\\ \boldsymbol{S}_{10} & \boldsymbol{0}\end{bmatrix}$ and $\boldsymbol{B}^{(K)} = \begin{bmatrix}\boldsymbol{S}_{00} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{11}\end{bmatrix}$. The generalized eigenvectors corresponding to r largest eigenvalues are included in $\boldsymbol{W}_K = \left[\boldsymbol{w}_1, \cdots, \boldsymbol{w}_r\right] \in \Re^{2m_1\times r}$. For the Kth mode $\mathcal{M}_K$, the dynamic cointegration matrix $\boldsymbol{W}_{e,K}$ and $\boldsymbol{W}_{f,K}$ are generated as the top and bottom halves of $\boldsymbol{W}_K$, namely, $\boldsymbol{W}_K = \begin{bmatrix}\boldsymbol{W}_{e,K}\\ \boldsymbol{W}_{f,K}\end{bmatrix}$.

2.3. Adaptive cointegration analysis

ACA [16] was introduced for online applications where the parameters can be adjusted to track the slow variation of the cointegration relationship. Here only the critical steps are outlined, and the further details can be found in [16].

For online applications, the mode index K is dropped to simplify notation. At time step $(t+1)$, a new sample $\boldsymbol{x}_{t+1}^0$ is collected and scaled as $\boldsymbol{x}_{t+1}$. Similarly, $\boldsymbol{x}_{t+1}$ is decomposed into $\boldsymbol{x}_{t+1} \longrightarrow \left\{\boldsymbol{x}_{0,t+1}\ \boldsymbol{x}_{1,t+1}\ \boldsymbol{x}_{2,t+1}\right\}$. The cointegration variables $\boldsymbol{x}_{1,t+1}$ are extracted and the matrix $\boldsymbol{X}_{1,t+1} = \begin{bmatrix}\boldsymbol{X}_{1,t}^T & \boldsymbol{x}_{1,t+1}^T\end{bmatrix}^T$ is constructed for ACA. As described in Appendix A in [16], the prediction errors $\tilde{\boldsymbol{E}}_{0,t+1}$ and $\tilde{\boldsymbol{E}}_{1,t+1}$ are updated recursively, $\boldsymbol{A}_{t+1}$ and $\boldsymbol{B}_{t+1}$ are calculated adaptively based on $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$. Subsequently, the objective of ACA is transformed into solving the generalized EVD problem below:

$$\boldsymbol{A}_{t+1}\boldsymbol{W}_{t+1} = \boldsymbol{B}_{t+1}\boldsymbol{W}_{t+1}\bar{\boldsymbol{\Lambda}}_{t+1} \tag{6}$$

which is settled by a standard EVD. The elements in the diagonal matrix $\bar{\boldsymbol{\Lambda}}_{t+1}$ are listed in descending order. $\boldsymbol{W}_{t+1}$ is the corresponding eigenmatrix and $\boldsymbol{W}_{t+1} = \begin{bmatrix}\boldsymbol{W}_{e,t+1}\\ \boldsymbol{W}_{f,t+1}\end{bmatrix}$.

2.4. AttPSFA for a single dynamic mode

PSFA was proposed [27] and applied to monitoring linear nonstationary processes [23]. The slowest features were

extracted and the monitoring statistics were constructed to distinguish both nominal operating points and dynamic behaviors. Besides, it could deal with measurement noise and missing data conveniently [23].

In this section, a nonlinear extension of PSFA (called AttPSFA) is investigated for nonlinear nonstationary processes, where attention mechanism is adopted to focus on the global and local important features. The original data are mapped to a high-dimensional feature space and then PSFA is adopted to extract the significant slowest features. For a single dynamic mode $\mathcal{M}_K$, the residue of primary signal through CA is then combined with $\boldsymbol{X}_{2,K}$, for the offline training of AttPSFA.

AttPSFA seeks to extract the slowest varying nonlinear latent features from the time-varying signal $\tilde{\boldsymbol{X}}_{2,K}$, constructed by

$$\tilde{\boldsymbol{X}}_{2,K} = \left[\boldsymbol{X}_{1,K}\boldsymbol{W}_{f,K}^{\perp} \quad \boldsymbol{X}_{2,K}\right] \tag{7}$$

where $\boldsymbol{W}_{f,K}^{\perp} = \boldsymbol{I} - \boldsymbol{W}_{f,K}\left(\boldsymbol{W}_{f,K}^{T}\boldsymbol{W}_{f,K}\right)^{-1}\boldsymbol{W}_{f,K}^{T}$, and $\boldsymbol{I}$ is the identity matrix with appropriate dimension. We refer to $\tilde{\boldsymbol{X}}_{2,K}$ as the secondary nonstationary signal.

To capture the significant global and local information, as well as to model the nonlinearity in secondary nonstationary signal $\tilde{\boldsymbol{X}}_{2,K}$, the data are mapped onto a high-dimensional feature space based on the attention mechanism. Consider an attention function between $\tilde{\boldsymbol{x}}_2$ and $\phi(\tilde{\boldsymbol{x}}_2)$:

$$\phi_j(\tilde{\boldsymbol{x}}_2) = \frac{\tilde{\boldsymbol{x}}_2^T \boldsymbol{c}_j}{d} \tag{8}$$

where $\phi(\tilde{\boldsymbol{x}}_2) = \{\phi_j(\tilde{\boldsymbol{x}}_2)\} \in \Re^M$, and M is predefined by the user. $\boldsymbol{C} = \{\boldsymbol{c}_j\}$, $j = 1, ...M$ are a set of M keys. $d > 0$ is a scaling hyperparameter.

Attention mapping is defined as

$$\texttt{Attention}(\tilde{\boldsymbol{x}}_2, \boldsymbol{C}, \boldsymbol{V}^{\dagger}) = \sum_{j=1}^{M} \texttt{softmax}(\tilde{\boldsymbol{x}}_2, \boldsymbol{C})_j v_j^{\dagger}$$

in which $v_j^{\dagger}$ is the element in $\boldsymbol{V}^{\dagger}$ and $\boldsymbol{V}^{\dagger}$ is pseudo-inverse of $\boldsymbol{V}$, and $\boldsymbol{V}$ would be explained in (10). Besides,

$$\texttt{softmax}(\tilde{\boldsymbol{x}}_2, \boldsymbol{C})_j = \frac{\exp(\phi_j(\tilde{\boldsymbol{x}}_2))}{\sum_{j=1}^{M}\exp(\phi_j(\tilde{\boldsymbol{x}}_2))} \tag{9}$$

For convenience, the compatibility function of $\texttt{softmax}(\cdot)$ is denoted as ${}_0\boldsymbol{x}_2^{\phi}$ and the mapped data matrix is ${}_0\boldsymbol{X}_{2,K}^{\phi}$. The mean and standard deviation of ${}_0\boldsymbol{X}_{2,K}^{\phi}$ are calculated and denoted as $\tilde{\boldsymbol{\mu}}_K^{\phi}$ and $\tilde{\boldsymbol{\Sigma}}_K^{\phi}$. Then, data ${}_0\boldsymbol{X}_{2,K}^{\phi}$ are normalized (zero mean and unit variance) and the processed data are labeled as $\boldsymbol{X}_{2,K}^{\phi}$. The initial keys $\boldsymbol{C}^K$ of the current mode $\mathcal{M}_K$ are determined using an online k-means clustering algorithm [28] based on $\boldsymbol{C}^{K-1}$ and $\tilde{\boldsymbol{X}}_{2,K}$.

Consider representing the time-varying observations, $\boldsymbol{X}_{2,K}^{\phi} = \{\boldsymbol{x}_i^{\phi}\} \in \Re^{N_K \times M}$ using a state-space model with a first-order Markov chain architecture [29].

$$\begin{cases} \boldsymbol{x}_i^{\phi} = \boldsymbol{V}\boldsymbol{y}_i + \boldsymbol{e}_i, & \boldsymbol{e}_i \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_x\right) \\ \boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{y}_{i-1} + \boldsymbol{\varepsilon}_i, & \boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right) \\ \boldsymbol{y}_1 = \boldsymbol{u}, & \boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_1\right) \end{cases} \tag{10}$$

where $\boldsymbol{Y}_K = \{\boldsymbol{y}_i\} \in \Re^{N_K \times p_2}$ contains the latent variables, $p_2 < M$. $\boldsymbol{\Lambda} = diag\left(\lambda_1, \ldots, \lambda_{p_2}\right)$, with the constraint $\boldsymbol{\Lambda}^2 + \boldsymbol{\Sigma} = \boldsymbol{I}$. The emission matrix is $\boldsymbol{V} \in \Re^{M \times p_2}$ and measurement noise variance is $\boldsymbol{\Sigma}_x = diag\left(\sigma_1^2, \cdots, \sigma_M^2\right)$. Let $\theta_x = \left\{\boldsymbol{V}, \boldsymbol{\Sigma}_x\right\}$, $\theta_y = \left\{\boldsymbol{\Sigma}_1, \boldsymbol{\Lambda}\right\}$, $\theta = \left\{\theta_x, \theta_y\right\}$. The joint distribution and the complete log likelihood function are [27]

$$P\left(\boldsymbol{X}_{2,K}^{\phi} | \boldsymbol{Y}_K\right) = P\left(\boldsymbol{y}_1\right)\prod_{i=2}^{N_K} P\left(\boldsymbol{y}_i | \boldsymbol{y}_{i-1}\right)\prod_{i=1}^{N_K} P\left(\boldsymbol{x}_i^{\phi} | \boldsymbol{y}_i\right) \tag{11}$$

$$\begin{aligned} \log\ P\left(\boldsymbol{X}_{2,K}^{\phi}, \boldsymbol{Y}_K | \theta\right) &= \sum_{i=1}^{N_K} \log\ P\left(\boldsymbol{x}_i^{\phi} | \boldsymbol{y}_i, \theta_x\right) \\ &+ \log\ P\left(\boldsymbol{y}_1 | \boldsymbol{\Sigma}_1\right) + \sum_{i=2}^{N_K} \log\ P\left(\boldsymbol{y}_i | \boldsymbol{y}_{i-1}, \boldsymbol{\Lambda}\right) \end{aligned} \tag{12}$$

respectively, which is optimized using the expectation maximization (EM) method [30].

3. Proposed ACA-RAttPSFA-EWC

This paper investigates an adaptive monitoring framework for multimode nonstationary processes, where a normal mode is identified automatically without abundant prior knowledge for online applications. When a new mode is detected by ACA, a small amount of data are collected to retrain the CA in Section 2.2 and AttPSFA-EWC models in Section 3.1. For online monitoring, the long-equilibrium features from primary nonstationary signals are extracted firstly and the corresponding ACA model parameters are updated recursively to track the slow variation of cointegration relationship in Section 2.3. The remaining nonstationary information is processed by the proposed RAttPSFA-EWC, the parameters of which are updated adaptively in Section 3.2. Then, the monitoring procedure is summarized in Section 3.3. Eventually, the proposed ACA-RAttPSFA-EWC is compared with several advanced approaches in Section 3.4.

3.1. AttPSFA–EWC for multiple modes

The objective of this work is to introduce an adaptive monitoring method with continual learning ability for sequential nonstationary modes. One main contribution of this paper is proposing AttPSFA-EWC for multimode nonlinear dynamic process monitoring that addresses the secondary nonstationary signals using EWC to combat catastrophic forgetting and EM to obtain optimization solution for offline

data. For online monitoring, the model parameters are updated adaptively to track the slow variations.

For online monitoring of a mode K, initially the model parameters are pre-trained offline and then the model parameters are updated adaptively. When a new mode is detected by ACA, a small amount of data $\boldsymbol{X}_K$ are collected for offline training. The significant knowledge learned from AttPSFA is preserved when a new mode arrives. The objective of AttPSFA-EWC is also automatically modified by adding an extra quadratic term to AttPSFA, which represents regularization according to the importance of model parameters, as estimated by EWC [20]. Similar to [17], the objective function of K existing modes is formally described by

$$\begin{aligned} J(\theta) = &\log\ P\left(\boldsymbol{X}_{2,K}^{\phi}, \boldsymbol{Y}_K|\theta\right) \\ &- J_{reg}\left(\boldsymbol{V}_{\mathcal{M}_{K-1}}, \boldsymbol{\Omega}^{V}_{\mathcal{M}_{K-1}}, \boldsymbol{\Lambda}_{\mathcal{M}_{K-1}}, \boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}}\right) \end{aligned} \tag{13}$$

subject to the AttPSFA model (10) and

$$\begin{aligned} &J_{reg}\left(\boldsymbol{V}_{\mathcal{M}_{K-1}}, \boldsymbol{\Omega}^{V}_{\mathcal{M}_{K-1}}, \boldsymbol{\Lambda}_{\mathcal{M}_{K-1}}, \boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}}\right) \\ =&-\gamma_{1,K}\left\|\boldsymbol{V}-\boldsymbol{V}_{\mathcal{M}_{K-1}}\right\|^2_{\boldsymbol{\Omega}^{V}_{\mathcal{M}_{K-1}}} \\ &-\gamma_{2,K}\sum_{i=1}^{p_2}\Omega^{\lambda}_{\mathcal{M}_{K-1},i}\left(\lambda_i-\lambda_{\mathcal{M}_{K-1},i}\right)^2 \end{aligned} \tag{14}$$

where $\boldsymbol{V}_{\mathcal{M}_{K-1}}$ and $\boldsymbol{\Lambda}_{\mathcal{M}_{K-1}}$ are the optimal parameters of last mode $\mathcal{M}_{K-1}$, $\boldsymbol{\Omega}^{V}_{\mathcal{M}_{K-1}}$ and $\boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}}$ are the corresponding importances, $\Omega^{\lambda}_{\mathcal{M}_{K-1},i}$ is the ith element of the diagonal matrix $\boldsymbol{\Omega}^{\Lambda}_{\mathcal{M}_{K-1}}$. $\gamma_{1,K}$ and $\gamma_{2,K}$ are the hyper-parameters and pre-defined by the user. The objective (13) is optimized by EM and the detailed procedure can be found in [17]. Using the Kalman filter, the final posterior mean and covariance of latent variables $\boldsymbol{Y}_K$ are denoted as $\boldsymbol{\mu}_K$ and $\boldsymbol{U}_K$ respectively. Correspondingly, the final solution of (13) is denoted as $\left\{\boldsymbol{V}_{\mathcal{M}_K}, \boldsymbol{\Sigma}_{x,\mathcal{M}_K}, \boldsymbol{\Lambda}_{\mathcal{M}_K}, \boldsymbol{F}^{V}_{\mathcal{M}_K}, \boldsymbol{F}^{\Lambda}_{\mathcal{M}_K}\right\}$. $\boldsymbol{F}^{V}_{\mathcal{M}_K}$ and $\boldsymbol{F}^{\Lambda}_{\mathcal{M}_K}$ are the Fisher information matrices with regard to $\boldsymbol{V}_{\mathcal{M}_K}$ and $\boldsymbol{\Lambda}_{\mathcal{M}_K}$ after the offline training procedure, respectively.

3.2. Online RAttPSFA–EWC updating

For online applications, the parameters of AttPSFA-EWC model are updated adaptively. For notational simplicity, it is assumed that the data $\boldsymbol{x}_i$ and corresponding slow features $\boldsymbol{y}_i$ start from $i=1$ at the beginning and end at $i=t$ for each mode.

3.2.1. Objective function design

At time step t, data $\boldsymbol{X}_t = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ are collected and divided into $\boldsymbol{X}_t = \left[\boldsymbol{X}_{0,t}\ \boldsymbol{X}_{1,t}\ \boldsymbol{X}_{2,t}\right]$. Features with common trends are extracted by ACA in Section 2.3 and parameters $\Theta_t^{ACA} = \{\tilde{\boldsymbol{E}}_{0,t}, \tilde{\boldsymbol{E}}_{1,t}, \boldsymbol{A}_t, \boldsymbol{B}_t, \boldsymbol{W}_{f,t}, \boldsymbol{W}_{e,t}\}$ are obtained. More detailed information can be found in [16].

The remaining nonstationary information is constructed using $\tilde{\boldsymbol{X}}_{2,t} = \left[\boldsymbol{X}_{1,t}\boldsymbol{W}^{\perp}_{f,t}\ \boldsymbol{X}_{2,t}\right]$, where $\boldsymbol{W}^{\perp}_{f,t} = \boldsymbol{I} - \boldsymbol{W}_{f,t}\left(\boldsymbol{W}^{T}_{f,t}\boldsymbol{W}_{f,t}\right)^{-1}\boldsymbol{W}^{T}_{f,t}$. The mapped data are calculated by (9) and denoted as $\boldsymbol{X}^{\phi}_{2,t}$. For RAttPSFA-EWC, the expectation of complete likelihood is designed as

$$\hat{Q}_t = \sum_{i=1}^{t}\mathbb{E}\left[\log P\left(\boldsymbol{x}^{\phi}_{2,i}, \boldsymbol{y}_i|\theta\right)\right] \tag{15}$$

Based on (10)–(13), $\hat{Q}_t$ is described as

$$\begin{aligned} &\hat{Q}_t(\boldsymbol{\Lambda}, \boldsymbol{V}, \boldsymbol{\Sigma}_x) \\ =&-\frac{t}{2}\log\left|\boldsymbol{\Sigma}_x\right| - \frac{1}{2}tr\left(\boldsymbol{D}_t\boldsymbol{V}^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\right) - \frac{1}{2}tr\left(\boldsymbol{H}_t\boldsymbol{\Sigma}_x^{-1}\right) \\ &+tr\left(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\boldsymbol{L}_t\right) - \frac{t-1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}tr\left(\boldsymbol{E}_t\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right) \\ &-\frac{1}{2}tr\left(\boldsymbol{F}_t\boldsymbol{\Sigma}^{-1}\right) + tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{G}_t\right) \end{aligned} \tag{16}$$

where $\boldsymbol{D}_t = \sum_{i=1}^{t}\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_i^T|\boldsymbol{X}^{\phi}_{2,t}\right]$, $\boldsymbol{E}_t = \sum_{i=2}^{t}\mathbb{E}\left[\boldsymbol{y}_{i-1}\boldsymbol{y}_{i-1}^T|\boldsymbol{X}^{\phi}_{2,t}\right]$, $\boldsymbol{F}_t = \sum_{i=2}^{t}\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_i^T|\boldsymbol{X}^{\phi}_{2,t}\right]$, $\boldsymbol{G}_t = \sum_{i=2}^{t}\mathbb{E}\left[\boldsymbol{y}_i\boldsymbol{y}_{i-1}^T|\boldsymbol{X}^{\phi}_{2,t}\right]$, $\boldsymbol{H}_t = \sum_{i=1}^{t}\mathbb{E}\left[\boldsymbol{x}^{\phi}_{2,i}(\boldsymbol{x}^{\phi}_{2,i})^T\right]$, $\boldsymbol{L}_t = \sum_{i=1}^{t}\mathbb{E}\left[\boldsymbol{y}_i|\boldsymbol{X}^{\phi}_{2,t}\right]\{\boldsymbol{x}^{\phi}_{2,i}\}^T$. The optimal solution of (16) is denoted as $\theta_t = \left\{\boldsymbol{V}_t, \boldsymbol{\Sigma}_{x,t}, \boldsymbol{\Lambda}_t\right\}$.

At time step $(t+1)$, a normal sample $\boldsymbol{x}_{t+1}$ is collected and divided into $\boldsymbol{x}_{t+1} = \left[\boldsymbol{x}_{0,t+1}\ \boldsymbol{x}_{1,t+1}\ \boldsymbol{x}_{2,t+1}\right]$. The ACA parameters $\Theta_{t+1}^{ACA} = \{\tilde{\boldsymbol{E}}_{0,t+1}, \tilde{\boldsymbol{E}}_{1,t+1}, \boldsymbol{A}_{t+1}, \boldsymbol{B}_{t+1}, \boldsymbol{W}_{f,t+1}, \boldsymbol{W}_{e,t+1}\}$ are updated based on Θ_t^{ACA} and $\boldsymbol{x}_{1,t+1}$. Detailed calculation procedure can be found in [16].

The second nonstationary signal is constructed by $\tilde{\boldsymbol{x}}_{2,t+1} = \left[\boldsymbol{x}_{1,t+1}\boldsymbol{W}^{\perp}_{f,t+1}\ \ \boldsymbol{x}_{2,t+1}\right]$ and RAttPSF-EWC parameters are updated based on $\tilde{\boldsymbol{x}}_{2,t+1}$. The keys $\boldsymbol{C}_{t+1}$ are updated based on $\boldsymbol{C}_t$ and $\tilde{\boldsymbol{x}}_{2,t+1}$ according to online k-means clustering algorithm [28]. Subsequently, data $\boldsymbol{x}^{\phi}_{2,t+1}$ are acquired by (9) and $\boldsymbol{X}^{\phi}_{2,t+1} = \left[\boldsymbol{X}^{\phi}_{2,t}; \boldsymbol{x}^{\phi}_{2,t+1}\right]$. Here, the expectation of complete likelihood is calculated recursively as follows [31]:

$$\hat{Q}_{t+1}(\theta) = \hat{Q}_t(\theta) + \gamma_{t+1}\left(\mathbb{E}_{\hat{\theta}_t}\left[\log P\left(\boldsymbol{x}^{\phi}_{2,t+1}, \boldsymbol{y}_{t+1}, \theta\right)\right] - \hat{Q}_t(\theta)\right) \tag{17}$$

and $\hat{\theta}_{t+1} = \arg\max\hat{Q}_{t+1}$, γ_{t+1} is the forgetting factor.

Substituting (10) into (17), we get

$$\begin{aligned} J =&\hat{Q}_{t+1}(\boldsymbol{\Lambda}, \boldsymbol{V}, \boldsymbol{\Sigma}_x) \\ =&-\frac{1}{2}tr\left(\boldsymbol{D}_{t+1}\boldsymbol{V}^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\right) - \frac{1}{2}tr\left(\boldsymbol{H}_{t+1}\boldsymbol{\Sigma}_x^{-1}\right) \\ &-\frac{\bar{\gamma}_{t+1}}{2}\log\left|\boldsymbol{\Sigma}_x\right| + tr\left(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\boldsymbol{L}_{t+1}\right) - \frac{\hat{\gamma}_{t+1}}{2}\log|\boldsymbol{\Sigma}| \\ &-\frac{1}{2}tr\left(\boldsymbol{E}_{t+1}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right) - \frac{1}{2}tr\left(\boldsymbol{F}_{t+1}\boldsymbol{\Sigma}^{-1}\right) \\ &+tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{G}_{t+1}\right) \end{aligned} \tag{18}$$

where $\bar{\gamma}_{t+1} = t(1-\gamma_{t+1}) + \gamma_{t+1}, \hat{\gamma}_{t+1} = (t-1)(1-\gamma_{t+1}) + \gamma_{t+1}$ and

$$\begin{cases} \boldsymbol{D}_{t+1} = \left(1-\gamma_{t+1}\right)\boldsymbol{D}_t + \gamma_{t+1}\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^T|\boldsymbol{X}_{2,t+1}^{\phi}\right] \\ \boldsymbol{E}_{t+1} = \left(1-\gamma_{t+1}\right)\boldsymbol{E}_t + \gamma_{t+1}\mathbb{E}\left[\boldsymbol{y}_t\boldsymbol{y}_t|\boldsymbol{X}_{2,t+1}^{\phi}\right] \\ \boldsymbol{F}_{t+1} = \left(1-\gamma_{t+1}\right)\boldsymbol{F}_t + \gamma_{t+1}\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^T|\boldsymbol{X}_{2,t+1}^{\phi}\right] \\ \boldsymbol{G}_{t+1} = \left(1-\gamma_{t+1}\right)\boldsymbol{G}_t + \gamma_{t+1}\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_t^T|\boldsymbol{X}_{2,t+1}^{\phi}\right] \\ \boldsymbol{H}_{t+1} = \left(1-\gamma_{t+1}\right)\boldsymbol{H}_t + \gamma_{t+1}\mathbb{E}\left[\boldsymbol{x}_{2,t+1}\boldsymbol{x}_{2,t+1}^T\right] \\ \boldsymbol{L}_{t+1} = \left(1-\gamma_{t+1}\right)\boldsymbol{L}_t + \gamma_{t+1}\mathbb{E}\left[\boldsymbol{y}_{t+1}|\boldsymbol{X}_{2,t+1}^{\phi}\right]\{\boldsymbol{x}_{2,t+1}^{\phi}\}^T \end{cases}$$

3.2.2. *Solution of objective* (18)

For the proposed method, recursive expectation maximization (REM) [31] is used to optimize the objective (18) for every time instant and obtain the parameters $\theta_{t+1} = \left\{\boldsymbol{V}_{t+1}, \boldsymbol{\Sigma}_{x,t+1}, \boldsymbol{\Lambda}_{t+1}\right\}$. In E-steps, three sufficient statistics $\mathbb{E}\left[\boldsymbol{y}_{t+1}|\boldsymbol{X}_{2,t+1}^{\phi}\right], \mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_t^T|\boldsymbol{X}_{2,t+1}^{\phi}\right]$ and $\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^T|\boldsymbol{X}_{2,t+1}^{\phi}\right]$ are calculated and would be used for M-steps. It contains forward recursion and backward recursion, as summarized in Algorithm 1. In the forward recursion, the posterior distribution $p\left(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{2,1}^{\phi}, \ldots, \boldsymbol{x}_{2,t+1}^{\phi}\right) = \mathcal{N}\left(\boldsymbol{\mu}_{t+1}, \boldsymbol{U}_{t+1}\right)$, is realized by Kalman filter. In the backward recursion, the marginal posterior distribution is calculated by Rauch-Tung-Striebel (RTS) smoother. The initial settings of REM are $\left\{\boldsymbol{V}_{\mathcal{M}_K}, \boldsymbol{\Sigma}_{x,\mathcal{M}_K}, \boldsymbol{\Lambda}_{\mathcal{M}_K}, \boldsymbol{\mu}_K, \boldsymbol{U}_K\right\}$ for the mode $\mathcal{M}_K$.

Algorithm 1 Updating statistics of the E-steps at $(t+1)$th sampling instant recursively

Inputs: $\tilde{\boldsymbol{x}}_{2,t+1}^{\phi}, \boldsymbol{U}_t, \boldsymbol{\mu}_t, \boldsymbol{V}, \boldsymbol{\Sigma}_x, \boldsymbol{\Lambda}$

Outputs: $\mathbb{E}\left[\boldsymbol{y}_{t+1}|\boldsymbol{X}_{2,t+1}^{\phi}\right]$, $\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_t^T|\boldsymbol{X}_{2,t+1}^{\phi}\right]$, $\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^T|\boldsymbol{X}_{2,t+1}^{\phi}\right]$, $\boldsymbol{U}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{P}_t, \boldsymbol{K}_{t+1}$

1: Forward steps by the Kalman filter:
   a) Calculate the prior covariance: $\boldsymbol{P}_t = \boldsymbol{\Lambda}\left(\boldsymbol{U}_t - \boldsymbol{I}\right)\boldsymbol{\Lambda}^T + \boldsymbol{I}$
   b) Calculate the Kalman gain: $\boldsymbol{K}_{t+1} = \boldsymbol{P}_t\boldsymbol{V}^T\left(\boldsymbol{V}\boldsymbol{P}_t\boldsymbol{V}^T + \boldsymbol{\Sigma}_x\right)^{-1}$
   c) Update the mean: $\boldsymbol{\mu}_{t+1} = \boldsymbol{\Lambda}\boldsymbol{\mu}_t + \boldsymbol{K}_{t+1}\left(\tilde{\boldsymbol{x}}_{2,t+1}^{\phi} - \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{\mu}_t\right)$
   d) Calculate the posterior covariance: $\boldsymbol{U}_{t+1} = \left(\boldsymbol{I} - \boldsymbol{K}_{t+1}\boldsymbol{V}\right)\boldsymbol{P}_t$
2: Backward steps by the RTS smoother
   a) Initialize $\hat{\boldsymbol{\mu}}_{t+1} = \boldsymbol{\mu}_{t+1}, \hat{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_{t+1}$
   b) Gain: $\boldsymbol{J}_t = \boldsymbol{U}_t\boldsymbol{\Lambda}^T\boldsymbol{P}_t^{-1}$
   c) Mean: $\hat{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t + \boldsymbol{J}_t\left(\hat{\boldsymbol{\mu}}_{t+1} - \boldsymbol{\Lambda}\boldsymbol{\mu}_t\right)$
3: Calculate the sufficient statistics
   a) $\mathbb{E}\left[\boldsymbol{y}_{t+1}|\boldsymbol{X}_{2,t+1}^{\phi}\right] = \hat{\boldsymbol{\mu}}_{t+1}$
   b) $\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_t^T|\boldsymbol{X}_{2,t+1}^{\phi}\right] = \boldsymbol{J}_t\hat{\boldsymbol{U}}_{t+1} + \hat{\boldsymbol{\mu}}_{t+1}\hat{\boldsymbol{\mu}}_t^T$
   c) $\mathbb{E}\left[\boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^T|\boldsymbol{X}_{2,t+1}^{\phi}\right] = \hat{\boldsymbol{U}}_{t+1} + \hat{\boldsymbol{\mu}}_{t+1}\hat{\boldsymbol{\mu}}_{t+1}^T$

In M-steps, the critical parameters $\left\{\boldsymbol{V}, \boldsymbol{\Sigma}_x, \boldsymbol{\Lambda}\right\}$ are optimized alternatively. With regard to $\boldsymbol{V}$,

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{V}) = &-\frac{t(1-\gamma_{t+1})+\gamma_{t+1}}{2}\log\left|\boldsymbol{\Sigma}_x\right| - \frac{1}{2}tr\left(\boldsymbol{H}_{t+1}\boldsymbol{\Sigma}_x^{-1}\right) \\ &- \frac{1}{2}tr\left(\boldsymbol{D}_{t+1}\boldsymbol{V}^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\right) + tr\left(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{V}\boldsymbol{L}_{t+1}\right) \end{aligned} \tag{19}$$

Let the gradient be zero, and we get

$$\boldsymbol{V} = \boldsymbol{L}_{t+1}^T\boldsymbol{D}_{t+1}^{-1} \tag{20}$$

With regard to $\boldsymbol{\Sigma}_x = diag\left(\sigma_1^2, \cdots, \sigma_M^2\right)$,

$$\begin{aligned} J(\boldsymbol{\Sigma}_x) = \sum_{j=1}^{M}\frac{1}{\sigma_j^2}\Big(&-\frac{1}{2}\left\{\boldsymbol{V}\boldsymbol{D}_{t+1}\boldsymbol{V}^T\right\}_{jj} \\ &-\frac{1}{2}\left\{\boldsymbol{H}_{t+1}\right\}_{jj} + \left\{\boldsymbol{V}\boldsymbol{L}_{t+1}\right\}_{jj}\Big) - \frac{\bar{\gamma}_{t+1}}{2}\sum_{j=1}^{M}\log\sigma_j^2 \end{aligned}$$

Let the gradient be zero, we can get

$$\sigma_j^2 = \frac{1}{\bar{\gamma}_{t+1}}\left(\boldsymbol{v}_j\boldsymbol{D}_{t+1}\boldsymbol{v}_j^T - 2\boldsymbol{v}_j\boldsymbol{l}_j + \left\{\boldsymbol{H}_{t+1}\right\}_{j,j}\right) \tag{21}$$

where $\boldsymbol{v}_j$ is the jth row of $\boldsymbol{V}$, $\boldsymbol{l}_j$ is the jth line of $\boldsymbol{L}_{t+1}$ and $j = 1, \ldots, M$.

With regard to $\boldsymbol{\Lambda}$, the objective is reformulated as

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{\Lambda}) = &-\frac{\hat{\gamma}_{t+1}}{2}\sum_{k=1}^{p_2}\log\left(1-\lambda_k^2\right) - \frac{1}{2}\sum_{k=1}^{p_2}\left\{\boldsymbol{E}_{t+1}\right\}_{k,k}\frac{\lambda_k^2}{1-\lambda_k^2} \\ &-\frac{1}{2}\sum_{k=1}^{p_2}\left\{\boldsymbol{F}_{t+1}\right\}_{k,k}\frac{1}{1-\lambda_k^2} + \sum_{k=1}^{p_2}\left\{\boldsymbol{G}_{t+1}\right\}_{k,k}\frac{\lambda_k}{1-\lambda_k^2} \end{aligned}$$

where $\hat{\gamma}_{t+1} = (t-1)(1-\gamma_{t+1}) + \gamma_{t+1}$. Let the gradient be zero, we can get

$$\begin{aligned} &\hat{\gamma}_{t+1}\lambda_k^3 + \left\{\boldsymbol{G}_{t+1}\right\}_{k,k}\lambda_k^2 - \Big(\hat{\gamma}_{t+1} + \left\{\boldsymbol{E}_{t+1}\right\}_{k,k} \\ &+ \left\{\boldsymbol{F}_{t+1}\right\}_{k,k}\Big)\lambda_k + \left\{\boldsymbol{G}_{t+1}\right\}_{k,k} = 0 \end{aligned} \tag{22}$$

The solution of (22) is acquired analytically and $0 < \lambda_k < 1$, $\boldsymbol{\Lambda} = diag\left(\lambda_1, \ldots, \lambda_{p_2}\right)$. Repeat the E-steps and M-steps until convergence, then we can get the optimal parameters $\{\boldsymbol{V}_{t+1}, \boldsymbol{\Sigma}_{x,t+1}, \boldsymbol{\Lambda}_{t+1}\}$.

3.2.3. *Calculating importance measure*

The Fisher information matrix (FIM) is calculated by the covariance of the gradient of the log likelihood function at the local optimum [17]. At time step $(t+1)$, the gradient about $\boldsymbol{V}$ is

$$\begin{aligned} &\nabla_{\boldsymbol{V}}\log\ P\left(\boldsymbol{x}_{2,t+1}^{\phi}, \boldsymbol{y}_{t+1}|\theta_{t+1}\right) \\ =&\boldsymbol{\Sigma}_{x,t+1}^{-1}\left(\boldsymbol{V}_{t+1}\boldsymbol{y}_{t+1} - \boldsymbol{x}_{2,t+1}^{\phi}\right)\boldsymbol{y}_{t+1}^T \end{aligned} \tag{23}$$

Algorithm 2 Offline training of the proposed method

Inputs: Data $\boldsymbol{X}_K$, keys $\boldsymbol{C}^{K-1}$, AttPSFA-EWC parameters of mode $\mathcal{M}_{K-1}$ $\{\boldsymbol{V}_{\mathcal{M}_{K-1}}, \boldsymbol{\Lambda}_{\mathcal{M}_{K-1}}, \boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}}, \boldsymbol{\Omega}^\Lambda_{\mathcal{M}_{K-1}}\}$

Outputs: Keys $\boldsymbol{C}^K$, CA parameters $\left\{\boldsymbol{A}^{(K)}, \boldsymbol{B}^{(K)}\right\}$, AttPSFA-EWC parameters $\left\{\boldsymbol{V}_{\mathcal{M}_K}, \boldsymbol{\Sigma}_{x,\mathcal{M}_K}, \boldsymbol{\Lambda}_{\mathcal{M}_K}, \boldsymbol{F}^V_{\mathcal{M}_K}, \boldsymbol{F}^\Lambda_{\mathcal{M}_K}\right\}$, EM parameters $\left\{\boldsymbol{\mu}_K, \boldsymbol{U}_K\right\}$

1: For the Kth mode, divide the data $\boldsymbol{X}_K$ into three parts according to correlation analysis and expert experience, namely, $\boldsymbol{X}_K = \left[\boldsymbol{X}_{0,K} \ \boldsymbol{X}_{1,K} \ \boldsymbol{X}_{2,K}\right]$;

2: Perform traditional CA on $\boldsymbol{X}_{1,K}$ and get the initial parameters $\{\boldsymbol{A}^{(K)}, \boldsymbol{B}^{(K)}, \boldsymbol{W}_{f,K}, \boldsymbol{W}_{e,K}\}$:

a) Calculate two prediction errors via least squares;

b) Compute $\boldsymbol{A}^{(K)}$ and $\boldsymbol{B}^{(K)}$, obtain $\boldsymbol{W}_{f,K}$ and $\boldsymbol{W}_{e,K}$ via solving (5);

3: Construct $\tilde{\boldsymbol{X}}_{2,K}$ and perform AttPSFA-EWC on $\tilde{\boldsymbol{X}}_{2,K}$:

a) According to online k-means clustering, calculate the keys $\boldsymbol{C}^K$ based on $\boldsymbol{C}^{K-1}$ and $\tilde{\boldsymbol{X}}_{2,K}$;

b) Map data $\tilde{\boldsymbol{X}}_{2,K}$ to a high-dimensional space via (9) based on $\boldsymbol{C}^K$, and denoted as ${}_0\boldsymbol{X}^\phi_{2,K}$. The mean and standard derivation are denoted as $\tilde{\boldsymbol{\mu}}^\phi_K$ and $\tilde{\boldsymbol{\Sigma}}^\phi_K$, and the processed data are labeled as $\boldsymbol{X}^\phi_{2,K}$.

c) Perform PSFA-EWC on $\boldsymbol{X}^\phi_{2,K}$ [17]:

i) Calculate sufficient statistics via Kalman filter and RTS;

ii) Maximize (12) to get $\left\{\boldsymbol{V}_{\mathcal{M}_K}, \boldsymbol{\Sigma}_{x,\mathcal{M}_K}, \boldsymbol{\Lambda}_{\mathcal{M}_K}\right\}$;

iii) Calculate FIM $\boldsymbol{F}^V_{\mathcal{M}_K}$ and $\boldsymbol{F}^\Lambda_{\mathcal{M}_K}$;

iv) The posterior mean and covariance of latent variables $\boldsymbol{Y}$ are $\boldsymbol{\mu}_K$ and $\boldsymbol{U}_K$.

Then, define

$$\begin{aligned}\boldsymbol{f}^V_{t+1} =& \boldsymbol{\Sigma}^{-1}_{x,t+1}\left(\boldsymbol{V}_{t+1}\boldsymbol{y}_{t+1} - \boldsymbol{x}^\phi_{2,t+1}\right)\boldsymbol{y}^T_{t+1}\boldsymbol{y}_{t+1} \\ & \left(\boldsymbol{V}_{t+1}\boldsymbol{y}_{t+1} - \boldsymbol{x}^\phi_{2,t+1}\right)^T\boldsymbol{\Sigma}^{-1}_{x,t+1}\end{aligned} \tag{24}$$

The FIM about $\boldsymbol{V}$ is updated as follows:

$$\boldsymbol{F}^V_{t+1} = \boldsymbol{F}^V_t + \boldsymbol{f}^V_{t+1} \tag{25}$$

Similarly, the gradient and FIM with respect to λ_k is

$$\begin{aligned}&\nabla_{\lambda_k}\log\ P\left(\boldsymbol{x}^\phi_{2,t+1}, \boldsymbol{y}_{t+1}|\theta_{t+1}\right) \\ =&\frac{-\lambda^3_k + y_{t+1,k}y_{t,k}\lambda^2_k + \left(1 - y^2_{t+1,k} - y^2_{t,k}\right)\lambda_k + y_{t+1,k}y_{t,k}}{\left(1-\lambda^2_k\right)^2} \\ \triangleq& g\left(y_{t+1,k}, y_{t,k}, \lambda_k\right)\end{aligned}$$

$$f^{\lambda_k}_{t+1} = g\left(y_{t+1,k}, y_{t,k}, \lambda_{t+1,k}\right)^2, k = 1, \cdots, p_2$$

and $\boldsymbol{f}^\Lambda_{t+1} = diag\left(f^{\lambda_1}_{t+1}, \cdots, f^{\lambda_{p_2}}_{t+1}\right)$. Then, the FIM about $\boldsymbol{\Lambda}$ is updated by

$$\boldsymbol{F}^\Lambda_{t+1} = \boldsymbol{F}^\Lambda_t + \boldsymbol{f}^\Lambda_{t+1} \tag{26}$$

3.3. Summary of the monitoring procedure

At time step $(t+1)$, a collected sample $\boldsymbol{x}^0_{t+1}$ is scaled as $\boldsymbol{x}_{t+1}$, and divided as $\boldsymbol{x}_{t+1} = \left[\begin{array}{ccc}\boldsymbol{x}_{0,t+1} & \boldsymbol{x}_{1,t+1} & \boldsymbol{x}_{2,t+1}\end{array}\right]$. $\boldsymbol{W}_{f,t}$ and $\boldsymbol{W}_{e,t}$ have already been obtained by ACA after time step t. Let $\hat{\boldsymbol{x}}_{1,t+1} = \left[\boldsymbol{x}_{1,t+1}\boldsymbol{W}_{f,t} \quad \boldsymbol{x}_{0,t+1}\right]$, T^2_f and T^2_e are designed to reflect the static and dynamic long-term equilibrium relationships, which would be used to identify the modes.

$$T^2_f = \hat{\boldsymbol{x}}_{1,t+1}\hat{\boldsymbol{x}}^T_{1,t+1} \tag{27}$$

$$T^2_e = \boldsymbol{e}_{0,t+1}\boldsymbol{W}_{e,t}\boldsymbol{W}^T_{e,t}{\boldsymbol{e}_{0,t+1}}^T \tag{28}$$

where the prediction error $\boldsymbol{e}_{0,t+1}$ is the last sample of $\tilde{\boldsymbol{E}}_{0,t+1}$.

The remaining information of ACA is constructed by $\hat{\boldsymbol{x}}_{2,t+1} = \left[\boldsymbol{x}_{1,t+1}\boldsymbol{W}^\perp_{f,t} \ \boldsymbol{x}_{2,t+1}\right]$. The mapped sample $\hat{\boldsymbol{x}}^\phi_{2,t+1}$ is calculated using (9) based on $\hat{\boldsymbol{x}}_{2,t+1}$ and $\boldsymbol{C}_t$. Similar to

Algorithm 3 Online monitoring of the proposed method

Inputs: Keys $\boldsymbol{C}^K$, CA parameters $\left\{\boldsymbol{A}^{(K)}, \boldsymbol{B}^{(K)}\right\}$, RAttPSFA-EWC parameters $\left\{\boldsymbol{V}_{\mathcal{M}_K}, \boldsymbol{\Sigma}_{x,\mathcal{M}_K}, \boldsymbol{\Lambda}_{\mathcal{M}_K}, \boldsymbol{\Omega}^V_{\mathcal{M}_K}, \boldsymbol{\Omega}^\Lambda_{\mathcal{M}_K}\right\}$, REM parameters $\left\{\boldsymbol{\mu}_K, \boldsymbol{U}_K\right\}$

1: Initialize $t = N_K$, $\boldsymbol{C}_t = \boldsymbol{C}^K$, $\boldsymbol{A}_t = \boldsymbol{A}^{(K)}$, $\boldsymbol{B}_t = \boldsymbol{B}^{(K)}$, $\boldsymbol{V}_t = \boldsymbol{V}_{\mathcal{M}_K}$, $\boldsymbol{\Sigma}_{x,t} = \boldsymbol{\Sigma}_{x,\mathcal{M}_K}$, $\boldsymbol{\Lambda}_t = \boldsymbol{\Lambda}_{\mathcal{M}_K}$, $\boldsymbol{F}^V_t = N_K\boldsymbol{F}^V_{\mathcal{M}_K}$, $\boldsymbol{F}^\Lambda_t = N_K\boldsymbol{F}^\Lambda_{\mathcal{M}_K}$, $\boldsymbol{\mu}_t = \boldsymbol{\mu}_K$, $\boldsymbol{U}_t = \boldsymbol{U}_K$;

2: Collect a sample $\boldsymbol{x}_{t+1}$ and divide the sample into three blocks, namely, $\boldsymbol{x}_{t+1} = \left[\begin{array}{ccc}\boldsymbol{x}_{0,t+1} & \boldsymbol{x}_{1,t+1} & \boldsymbol{x}_{2,t+1}\end{array}\right]$;

3: Construct $\hat{\boldsymbol{x}}_{1,t+1} = \left[\boldsymbol{x}_{1,t+1}\boldsymbol{W}_{f,t} \ \boldsymbol{x}_{0,t+1}\right]$ and $\hat{\boldsymbol{x}}_{2,t+1} = \left[\boldsymbol{x}_{1,t+1}\boldsymbol{W}^\perp_{f,t} \ \boldsymbol{x}_{2,t+1}\right]$, and calculate $\hat{\boldsymbol{x}}^\phi_{2,t+1}$ based on $\hat{\boldsymbol{x}}_{2,t+1}$ and $\boldsymbol{C}_t$;

4: Calculate test statistics via (27)–(31) and judge the operating conditions using monitoring rules detailed in Section 3.3:

a) Normal, go to step 5 and update ACA-RAttPFSA-EWC parameters;

b) The mode is switched normally. Let $\boldsymbol{C}_K = \boldsymbol{C}_t$, $\boldsymbol{\Lambda}_{\mathcal{M}_K} = \boldsymbol{\Lambda}_t$, $\boldsymbol{V}_{\mathcal{M}_K} = \boldsymbol{V}_t$, $\boldsymbol{\Omega}^V_{\mathcal{M}_K} = \boldsymbol{\Omega}^V_{\mathcal{M}_{K-1}} + \frac{1}{t}\boldsymbol{F}^V_t$, $\boldsymbol{\Omega}^\Lambda_{\mathcal{M}_K} = \boldsymbol{\Omega}^\Lambda_{\mathcal{M}_{K-1}} + \frac{1}{t}\boldsymbol{F}^\Lambda_t$, $K = K+1$. Collect normal data $\boldsymbol{X}^0_K \in \Re^{n_0\times m}$, call Algorithm 2; Return to step 1.

c) Faulty, alarm is triggered.

5: Update the thresholds by RKDE;

6: Calculate two prediction errors $\boldsymbol{A}_{t+1}$ and $\boldsymbol{B}_{t+1}$, and get the parameters $\{\boldsymbol{W}_{f,t+1}, \boldsymbol{W}_{e,t+1}\}$ by solving (6);

7: Construct $\tilde{\boldsymbol{x}}_{2,t+1} = \left[\boldsymbol{x}_{1,t+1}\boldsymbol{W}^\perp_{f,t+1} \ \boldsymbol{x}_{2,t+1}\right]$, perform RAttPSFA-EWC on $\tilde{\boldsymbol{x}}_{2,t+1}$:

a) According to online k-means clustering, update the keys $\boldsymbol{C}_{t+1}$ based on $\tilde{\boldsymbol{x}}_{2,t+1}$ and $\boldsymbol{C}_t$;

b) The high-dimensional sample ${}_0\boldsymbol{x}^\phi_{2,t+1}$ is obtained using (9) based on $\boldsymbol{C}_{t+1}$ and $\tilde{\boldsymbol{x}}_{2,t+1}$, the preprocessed sample is denoted as $\boldsymbol{x}^\phi_{2,t+1}$;

c) Optimize the objective (17) based on REM:

i) Calculate three sufficient statistics using Algorithm 1;

ii) Update parameters using (20)–(22);

iii) Return to step i) until convergence;

d) Update the FIMs by (25) and (26);

8: Move to the next time step $t+1$, return to step 2.

Table 1
Descriptions of different monitoring subspaces

	Subspace	Dimension	Statistics	Description
ACA	Static subspace	m_1	T_f^2	Monitor the static long-term equilibrium relation
	Dynamic subspace	m_1	T_e^2	Monitor the dynamic long-term equilibrium relation
RAttPSFA-EWC	Static subspace	p_2	T^2	Monitor the static slow variations
	Static subspace	M	SPE	Monitor the prediction error
	Dynamic subspace	p_2	S^2	Monitor the dynamic slow variations

[17], three monitoring statistics are designed to monitor the short-term dynamics. According to Kalman filter equation,

$$\boldsymbol{y}_{t+1} = \boldsymbol{\Lambda}_{t+1}\boldsymbol{y}_t + \boldsymbol{K}_{t+1}\left[\hat{\boldsymbol{x}}_{2,t+1}^{\phi} - \boldsymbol{V}_{t+1}\boldsymbol{\Lambda}_{t+1}\boldsymbol{y}_t\right]$$

where $\boldsymbol{K}_{t+1}$ is the Kalman gain and calculated in Algorithm 1. Then, the T^2 statistic is defined as

$$T^2 = \boldsymbol{y}_{t+1}^T\boldsymbol{y}_{t+1} \tag{29}$$

To design the SPE statistic, the bias between the true value and one-step prediction is calculated at time step $(t+1)$. The prediction error follows Gaussian distribution, namely

$$\boldsymbol{\varepsilon}_{t+1} = \hat{\boldsymbol{x}}_{2,t+1}^{\phi} - \boldsymbol{V}_{t+1}\boldsymbol{\Lambda}_{t+1}\boldsymbol{\mu}_t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Phi}_{t+1}\right)$$

where $\boldsymbol{\Phi}_{t+1} = \boldsymbol{V}_{t+1}\boldsymbol{\Lambda}_{t+1}\boldsymbol{P}_t\boldsymbol{\Lambda}_{t+1}^T\boldsymbol{V}_{t+1}^T + \boldsymbol{\Sigma}_{x,t+1} + \boldsymbol{V}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{V}_{t+1}^T$. $\boldsymbol{\mu}_t$ is the prior mean and $\boldsymbol{P}_t$ is the prior covariance, which are calculated by Algorithm 1. The SPE statistic is designed to characterize the noise and calculated by

$$SPE = \boldsymbol{\varepsilon}_{t+1}^T\boldsymbol{\Phi}_{t+1}^{-1}\boldsymbol{\varepsilon}_{t+1} \tag{30}$$

S^2 statistic is designed to monitor temporal dynamics[23].

$$S^2 = \dot{\boldsymbol{y}}_{t+1}^T\boldsymbol{\Xi}^{-1}\dot{\boldsymbol{y}}_{t+1} \tag{31}$$

where $\dot{\boldsymbol{y}}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{y}_t$, $\boldsymbol{\Xi} = 2\left(\boldsymbol{I}_{p_2} - \boldsymbol{\Lambda}_{t+1}\right)$ [23].

For the proposed ACA-RAttPSFA-EWC, data are decomposed into five subspaces and the corresponding statistics are designed to reflect the variations, as summarized in Table 1. The offline thresholds are estimated by kernel density estimation (KDE) and the online thresholds are updated by recursive KDE (RKDE) [16]. The offline training and online monitoring procedures are summarized in Algorithms 2 and 3, respectively. The comprehensive procedure of ACA-RAttPSFA-EWC is depicted in Figure 1. The major differences between ACA-RPCA-EWC and ACA-RAttPSFA-EWC are highlighted.

The automatic monitoring is enabled with Line 4 in Algorithm 2 which continuously assesses three conditions online: a) No fault and same mode: continue using online monitoring; (b) No fault with new mode being detected, move to next mode using Algorithm 1 for offline training, which then returns to Step 2 of Algorithm 2; and (c) fault is triggered with report. Specifically, the monitoring rules in Algorithm 3 are summarized as follows:

a) If all statistics are below their thresholds, the process operates normally in the same mode. The ACA-RAttPSFA-EWC parameters are updated adaptively;

b) If T_f^2 or T_e^2 is out of control, it indicates that the static or dynamic long-term equilibrium relationship between cointegration variables is broken. If T_e^2 returns to normal, process dynamics are still controlled and a new mode arrives, and then a few normal samples are collected to establish the initial monitoring model. Otherwise, a fault is detected and a fault alarm is triggered;

c) If T^2 or SPE is above the threshold, a steady deviation from the predefined operating modes occurs. If S^2 is beyond the threshold, it indicates that a potential anomaly may have occurred and the process needs to be checked carefully.

3.4. Discussion

RSFA [12], RCA [11], ACA-RPCA-EWC [16], PSFA-EWC [17] are adopted to compare with the proposed ACA-RAttPSFA-EWC. Aforementioned methods are based on SFA or CA, and deeply intertwined. RSFA, PSFA-EWC and ACA-RAttPSFA-EWC are built on the foundation of SFA and also share the virtues of SFA, which focus on the slow variations of dynamics. In addition, PSFA-EWC and the proposed method can deal with measurement noise and missing data owing to probabilistic interpretation, where EM is adopted to optimize the parameters. Except for PSFA-EWC, the parameters of four methods are updated recursively online and the mode is identified automatically.

Comprehensive comparison of five methods are summarized briefly in Table 2. Several critical characteristics are discussed deeply to reflect the performance.

a) *Mode identification*. Real fault, normal mode switching and nonstationarity may occur in multimode nonstationary processes, which can make data vary dynamically. RSFA and PSFA-EWC cannot identify the modes without human intervention, because it is difficult to judge the root cause of dynamic variations. In some practical situations, it is hard to obtain the mode information in advance. RCA, ACA-RPCA-EWC and ACA-RAttPSFA-EWC can identify modes automatically, and distinguish the mode switching and real faults. Since manipulated variables are considered in ACA-RPCA-EWC and ACA-RAttPSFA-EWC, they are robust to the mode misidentification caused by human parameter adjustment [16].

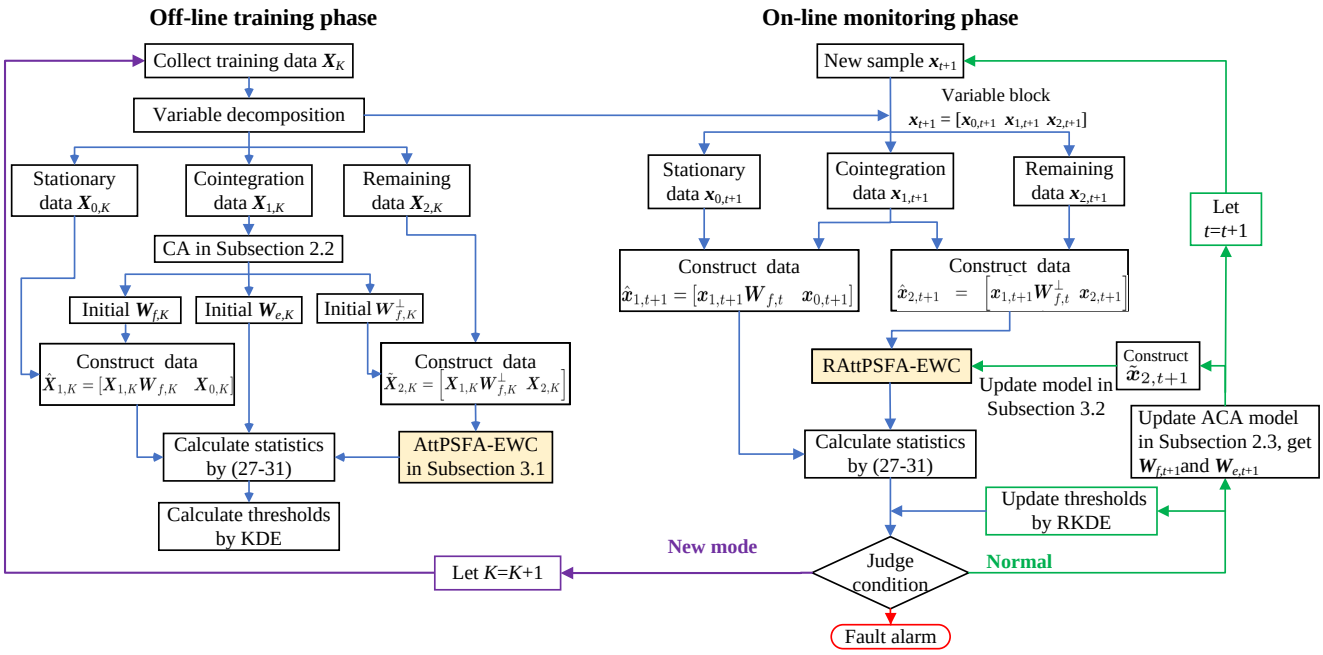


Figure 1: The flowchart of the proposed method

Table 2
Performance comparison of five methods in multimode nonstationary process monitoring

Methods	Intelligent mode identification	Model track accuracy	Nonlinearity	Dealing with uncertainty	Memory properties	Online real-time performance	Algorithm complexity
RSFA [12]	Poor	Poor	No	No	No	Poor	Low
RCA [11]	Good	Poor	No	No	No	Good	Low
PSFA-EWC [17]	Poor	Poor	No	Yes	Yes	Poor	Low
ACA-RPCA-EWC [16]	Excellent	Excellent	No	No	Yes	Excellent	Medium
ACA-RAttPSFA-EWC	Excellent	Excellent	Yes	Yes	Yes	Excellent	High

b) *Feature extraction.* ACA-RPCA-EWC and the proposed ACA-RAttPSFA-EWC method both extract significant features to monitor the long-term equilibrium relation. As illustrated in Table 1, the latter one further extracts dynamic and static features to monitor the slow variation and prediction error, while ACA-RPCA-EWC only considers static features after ACA. Although the procedure of AttPSFA-EWC and PSFA-EWC is similar, features with long-term equilibrium relation are neglected for PSFA-EWC. This characteristic is also applied to RSFA. Conversely, RCA merely extracts features with common trends and the remaining features are ignored.

c) *Nonlinearity and dealing with uncertainty*. Attention mechanism is adopted in the proposed ACA-RAttPSFA-EWC method, where data are mapped to a high-dimensional space to cope with nonlinearity and latent variables are extracted thereafter. The other four methods are applied to linear nonstationary processes. Moreover, PSFA-EWC and ACA-RAttPSFA-EWC use probabilistic interpretation to characterize uncertainty and EM is adopted to solve the optimization issue, which makes them also potentially deal with missing data. However, RSFA, RCA and ACA-RPCA-EWC fail to deal with uncertainty.

d) *Model track accuracy*. RSFA and RCA are designed for a single nonstationary mode and may fail to provide tracking performance for multimode nonstationary processes. PSFA-EWC requires a few representative data when a new mode arrives. Then, the model is trained and would be used without any updating for online monitoring. The performance may be decreased abruptly if the modes vary significantly. ACA-RPCA-EWC and ACA-RAttPSFA-EWC are mode-free monitoring methods, which could offer tracking accuracy owing to the comprehensive variable decomposition, adaptive updating and the previously consolidated knowledge. Furthermore, ACA-RAttPSFA-EWC may provide better performance than ACA-RPCA-EWC because dynamic characteristics are further extracted after ACA algorithm.

e) *Memory properties*. RSFA and RCA do not store the previously learned knowledge. PSFA-EWC calculates the FIM after the training procedure and would not be updated before a new mode arrives. For ACA-RPCA-EWC, the FIM of a certain mode is calculated at the end of each mode. For ACA-RAttPSFA-EWC, the FIM of each sample is calculated at each sampling instant and thus the FIM of a mode is obtained finally once

a new mode arrives. Theoretically, the FIM of ACA-RAttPSFA-EWC may contain more significant information. Thus, the memory characteristics may be better to other methods.

f) *Online real-time performance*. This paper mainly refers to the real-time monitoring performance when a new mode arrives. PSFA-EWC trains the parameters offline and then the model is used for online monitoring. When a new mode appears, a certain amount of data should be collected to retrain the model. During this period, the systems are monitored by an inaccurate model, which may influence the online real-time performance. When the mode switches, RSFA could not track the rapid and dynamic variations based on a few data, and thus the online monitoring would be unsatisfactory. RCA can extract rough long-term equilibrium features and the model is corrected based on the forthcoming data. ACA-RPCA-EWC and ACA-RAttPSFA-EWC provide a comprehensive monitoring framework and the significant features are extracted deeply, which are beneficial to establishing an accurate model based on limited data. Thus, their online monitoring performance would be excellent and optimal among five methods.

g) *Algorithm complexity*. This mainly refers to the online computational complexity. PSFA-EWC calculates the three statistics based on the trained parameters without updating, thus the online complexity may be the least. The complexity of RSFA and RCA comes in second place, because the parameter updating rules are intuitively explicit. Compared to RSFA and RCA, the complexity of ACA-RPCA-EWC is a little higher because more variables are considered and more parameters need to be updated. The complexity of ACA-RAttPSFA-EWC is particularly complicated because REM algorithm is adopted to optimize the objective function.

4. Experimental analysis

In this section, RSFA [12], RCA [11], ACA-RPCA-EWC [16], PSFA-EWC [17] are compared in a numerical case and a pulverizing system case. The offline training and online monitoring data are the same for all methods. Fault detection rates (FDRs), false alarm rates (FARs) and detection delay (DD) are used to evaluate the performance.

Note that the monitoring statistics of ACA-RAttPSFA-EWC and ACA-RPCA-EWC can be separated into tow parts, namely, the ACA and the remaining part. Specifically, ACA-RAttPSFA-EWC and ACA-RPCA-EWC adopt ACA to identify the mode and the corresponding ACA parameters are the same. Therefore, they share the same results of T_f^2 and T_e^2 in the following experiments. The monitoring indexes T_e^2 and T_f^2 of ACA are listed under ACA-RAttPSFA-EWC in Table 3. For ACA-RPCA-EWC, the remaining statistics T^2 and SPE are provided by RPCA-EWC separately. With regard to ACA-RAttPSFA-EWC, the rest monitoring statistics T^2, SPE and S^2 are calculated by RAttPSFA-EWC.

4.1. Numerical cases

Consider the following numerical data [16]:

$$\begin{cases} z_1 = a_1 t + b_1 + \varepsilon_1 \\ z_2 = a_2 t + b_2 + \varepsilon_2 \\ z_3 = a_3 t^2 + b_3 t + c_3 + \varepsilon_3 \\ z_4 = a_4 t^2 + b_4 t + c_4 + \varepsilon_4 \\ z_5 = a_5 + \varepsilon_5 \\ z_6 = a_6 e^{-t} + b_6 t + c_6 \sin t + d_6 + \varepsilon_6 \\ z_7 = a_7 e^{-t} + b_7 t^3 + c_7 \cos t + d_7 + \varepsilon_7 \end{cases}$$

where noise $\varepsilon_j \sim \mathcal{N}(0, 0.09)$, $j = 1, \ldots, 7$. The coefficients are shown as follows.

For mode $\mathcal{M}_1$:

$$\begin{cases} a_1 = 1.5, b_1 = 4; \\ a_2 = 1, b_2 = 2.5; \\ a_3 = -0.8, b_3 = 1.6, c_3 = 1; \\ a_4 = 0.6, b_4 = -1.2, c_4 = 2; \\ a_5 = 3; \\ a_6 = 0.4, b_6 = -0.1, c_6 = 0.2, d_6 = 0.8; \\ a_7 = 0.6, b_7 = 0.1, c_7 = 0.6, d_7 = 0.4; \end{cases} \tag{32}$$

For mode $\mathcal{M}_2$:

$$\begin{cases} a_1 = 1.5, b_1 = 3.5; \\ a_2 = 2, b_2 = 2; \\ a_3 = 0.4, b_3 = -0.8, c_3 = 2; \\ a_4 = -0.2, b_4 = 0.4, c_4 = 1.5; \\ a_5 = 2; \\ a_6 = 0.6, b_6 = -0.1, c_6 = 0.4, d_6 = 0.8; \\ a_7 = 0.6, b_7 = 0.3, c_7 = 0.4, d_7 = 0.4; \end{cases} \tag{33}$$

For mode $\mathcal{M}_3$:

$$\begin{cases} a_1 = 1.2, b_1 = 3; \\ a_2 = 2, b_2 = 2.5; \\ a_3 = 0.4, b_3 = -0.8, c_3 = 1; \\ a_4 = -0.3, b_4 = 0.6, c_4 = 1.5; \\ a_5 = 1.6; \\ a_6 = 0.4, b_6 = -0.1, c_6 = 0.3, d_6 = 0.6; \\ a_7 = 0.5, b_7 = 0.2, c_7 = 0.5, d_7 = 0.8; \end{cases} \tag{34}$$

Data from three successive modes are collected and 500 samples are generated from each mode. There are 1200 normal samples and the faulty data are generated as follows:

a) Case 1: z_1 is added 0.5 from the 201th sample in mode $\mathcal{M}_3$.

b) Case 2: z_6 is added 0.8 from the 201th sample in mode $\mathcal{M}_3$.

z_1 and z_2 are nonstationary and share the same trend. The same is true for z_3 and z_4. z_5 is stationary for each mode and would change when the mode switches. z_6 and z_7 change irregularly and dramatically. According to the prior knowledge, ADF test and correlation analysis, the variables

Table 3
Evaluation indices of the numerical case and the practical coal pulverizing case studies

Case number	Indices	RSFA			RCA			PSFA-EWC			ACA-RPCA-EWC		ACA-RAttPSFA-EWC				
		T^2	S^2	D^2	T^2	T_e^2	S^2	T^2	T_e^2	S^2	T^2	SPE	T_f^2	T_e^2	T^2	SPE	S^2
Case 1	FDRs	0	0	0.33	1.00	31.00	0	100	100	3.00	0	0	99.67	99.67	98.67	99.34	0
	FARs	3.91	7.36	1.09	2.73	14.55	2.73	1.36	4.45	5.18	5.09	9.18	10.73	3.45	8.45	8.73	2.00
	DD	300	300	0	14	5	300	0	0	0	300	300	1	1	1	300	300
Case 2	FDRs	0	0	0.33	99.67	31.67	0	87.00	100	6.33	100	100	0	0	99.67	99.67	2.33
	FARs	3.91	7.36	1.09	2.73	14.55	2.73	4.09	6.00	6.18	5.09	9.18	10.75	3.45	11.36	9.73	3.45
	DD	300	300	0	1	1	300	0	0	5	0	0	300	300	1	1	1
Case 3	FDRs	0	100	0.92	0	0	0	100	100	100	0	0	6.08	96.11	90.06	75.87	90.26
	FARs	40.65	99.93	1.52	0.52	6.02	0	97.97	95.97	95.43	2.07	1.38	1.15	1.97	9.47	8.19	1.29
	DD	3601	0	64	3601	3601	3601	-	-	-	3601	3601	3380	11	308	309	453
Case 4	FDRs	100	0	5.56	0	0	3.82	100	100	100	74.02	81.33	74.27	99.76	91.34	95.16	85.69
	FARs	89.55	10.66	0.85	0.03	0.21	3.92	96.95	95.55	95.10	2.33	4.03	1.36	1.77	2.87	2.09	1.03
	DD	0	2067	130	2067	2067	234	-	-	-	537	386	532	5	179	97	296
Case 5	FDRs	0	0	10.68	98.18	0	0.25	100	100	100	98.16	98.33	98.43	99.83	97.63	97.63	97.59
	FARs	0	3.17	0.89	0.04	0.20	2.98	98.57	94.76	94.02	0.61	0.06	0.73	1.40	6.36	3.65	0.59
	DD	4720	4720	18	86	4720	51	-	-	-	87	51	74	3	112	112	114

Table 4
Offline training time (s) of all algorithms

Methods	Case 1	Case 2	Case 3	Case 4	Case 5
RSFA	0.1881	0.0083	0.0409	0.0132	0.0163
RCA	0.1178	0.0095	0.0244	0.0161	0.0337
PSFA-EWC	112.04	142.9973	23512.5	19430.1	17239.26
ACA-RPCA-EWC	0.1752	0.0460	0.0723	0.0717	0.0656
ACA-RAttPSFA-EWC	5.6099	5.8662	156.5373	59.9634	135.2942

Table 5
Online testing time (s) of all algorithms

Methods	Case 1	Case 2	Case 3	Case 4	Case 5
RSFA	7.8775	7.5701	582.2558	482.75	512.6431
RCA	1.4266	1.2026	40.6655	25.4663	25.4418
PSFA-EWC	0.0174	0.0145	1.5378	0.6454	3.2370
ACA-RPCA-EWC	3.7974	3.3710	60.5120	50.2188	102.5709
ACA-RAttPSFA-EWC	226.2915	226.2252	4651.879	7405.276	9775.15

are decomposed into three blocks, i.e., $\boldsymbol{x}_1 = \left[z_1, z_2, z_3, z_4\right]$, $\boldsymbol{x}_0 = z_5$, and $\boldsymbol{x}_2 = \left[z_6, z_7\right]$. In this experiment, 100 normal samples are adopted to train the initial model and then sequential normal samples are utilized to update the parameters. When a new mode arrives, 20 normal samples are collected to retrain the model offline. The five methods share the same training and testing data.

The simulation results are summarized in Table 3. RSFA fails to monitor Case 1 and Case 2 because the FDRs of three statistics are close to 0. Although the FDR of RCA in Case 2 is 99.67%, the FARs of two cases are 14.55%. Therefore, RCA could not detect the faults in the two cases accurately. PSFA-EWC can monitor two cases accurately, where the FDRs are 100% and the FARs are lower than 6.2%. For ACA-RAttPSFA-EWC and ACA-RPCA-EWC, the FDRs of T_f^2 and T_e^2 are 99.67% in Case 1, which is in accordance with the fact that the cointegration relationship is broken owing to the faulty variable z_1. The FDRs of RAttPSFA-EWC approach 100%, while the FDRs of RPCA-EWC are 0. ACA-RAttPSFA-EWC and ACA-RPCA-EWC provide a similar monitoring accuracy in Case 2. Since the fault occurs in variable z_6 and the cointegration relationship remains the same, the FDRs of T_f^2 and T_e^2 are 0. RAttPSFA-EWC and RPCA-EWC can detect the fault accurately, where the FDRs are 99.67% and 100% respectively. Besides, the FARs are lower than 12%.

The offline training time and online testing time are listed in Tables 4 and 5, respectively. RSFA and RCA consume similar computational resources for the training procedure. ACA-RPCA-EWC takes the second place, followed by ACA-RAttPSFA-EWC. PSFA-EWC costs the most expensive computational resources, because the retraining procedure would be conducted when a new mode arrives and the optimization issue is settled by EM. For online applications, 1400 samples are utilized. RCA is the least computationally complicated, ACA-RPCA-EWC and RSFA come second. With regard to ACA-RAttPSFA-EWC, the testing time is 0.1616 second on average for each testing sample, which is accepted. Different from the aforementioned adaptive methods, there is no need to update the PSFA-EWC model parameters and thus the testing time is the least.

In conclusion, ACA-RAttPSFA-EWC and ACA-RPCA-EWC can identify the mode automatically and monitor two cases accurately. PSFA-EWC can also provide excellent monitoring performance, but the mode identification may require the prior knowledge for online applications. RSFA and RCA could not track the rapid variations in multimode nonstationary processes and the monitoring performance is unsatisfactory.

4.2. Pulverizing System Case

4.2.1. System description

The coal pulverizing system of a 1030-MW ultra-supercritical thermal power plant in China is adopted to illustrate the

Table 6
Data information of the coal pulverizing system

Case number	Coal type	Number of testing data	Fault instant	Fault cause
Case 3	Yinni—Menghun—Yinni	14120	10520	The temperature of the coal mill increases abnormally
Case 4	Aoemng—Aomei—Aomeng	9800	7734	Opening of the regulating baffle of primary air is abnormally large
Case 5	Waigou—Shenhun	14120	9401	There is oil leakage at the bearing of the rotary separator motor

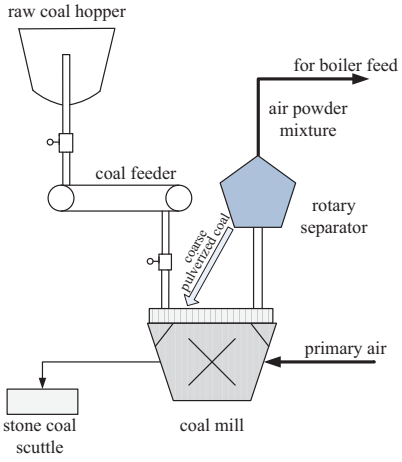


Figure 2: The structure of the coal pulverizing system

Table 7
Variable description of the coal pulverizing system

Description	Variable
Rotary separator speed	z_1
Coal mill seal air pressure	z_2
Differential pressure between seal air and grinding bowl	z_3
Upper and lower differential pressure of coal mill bowl	z_4
Instantaneous coal feeding capacity	z_5
Motor speed of coal feeder	z_6
Generator active power	z_7
Air powder mixture pressure	z_8
Cold primary air electric regulating baffle position feedback	z_9
Primary air flow	z_{10}
Primary air pressure	z_{11}
Coal mill current	z_{12}
Primary air temperature at the outlet of the air preheater	z_{13}, z_{14}
Coal feeder current	z_{15}
Outlet temperature	z_{16}
Planetary gear box input bearing temperature	z_{17}, z_{18}
Temperature of planetary gearbox bearings	$z_{19} \sim z_{22}$
Inlet air temperature of forced draft fan	z_{23}, z_{24}
Bearing temperature of rotary separator	z_{25}, z_{26}

effectiveness of the proposed method. It is an important auxiliary machinery and locates at the forefront of the power plant, which would influence the operating condition significantly. It aims to grind raw coal into coal powder with desired fineness and the temperature is also taken into consideration to guarantee the process safety and the combustion efficiency. The structure is depicted in Figure 2, including the coal feeder, rotary separator, coal mill, etc.

The process data of the large-scale thermal power plant are mainly affected by the coal and unit loading. Various types of coal may be fired according to the economic benefit and environmental requirement. For different types of coal, the process data may vary with the real-time load significantly. Thus, if one coal is regarded as a mode, the system is still nonstationary in each mode. According to the historical fault record and fault effects, two typical faults from the outlet temperature (Cases 3 and 4) and the rotary separator (Case 5) are considered in this paper. The detailed information of practical data is summarized in Table 6. The sampling interval is 20s. Note that it is difficult to estimate accurately when the coal changes, and therefore the accurate mode switching time is unavailable.

4.2.2. Data analysis

In this paper, 26 critical variables are selected and decomposed according to process mechanism, prior knowledge and expert experience. As listed in Table 7, the variables are divided into three blocks, i.e., $\boldsymbol{x}_1 = [z_1, z_2, \ldots, z_{12}]$, $\boldsymbol{x}_0 = [z_{13}, z_{14}, z_{15}, z_{16}]$, and $\boldsymbol{x}_2 = [z_{17}, z_{18}, \ldots, z_{26}]$. The variables in $\boldsymbol{x}_1$ are generally nonstationary in each mode

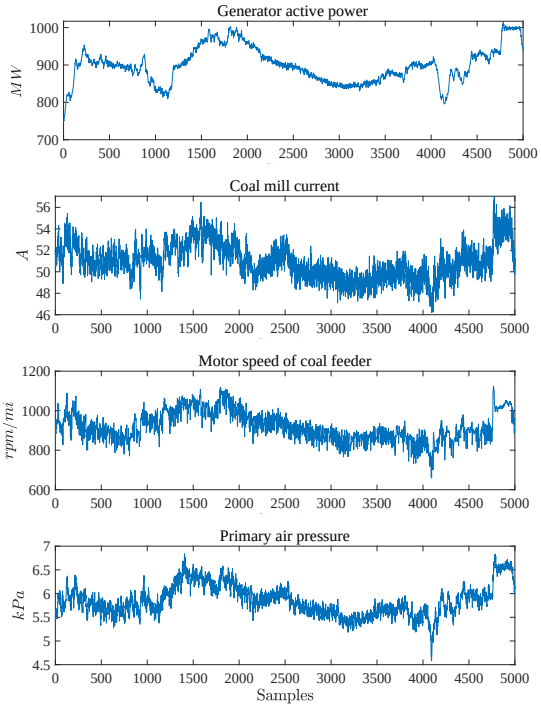
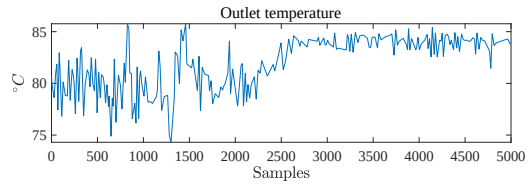
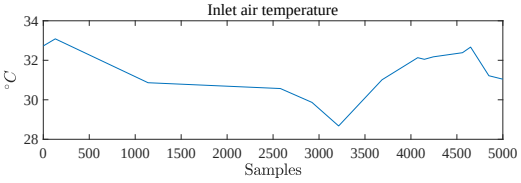


Figure 3: Partial cointegration variables in $\boldsymbol{x}_1$ of the practical system



(a) Partial stationary variable in $\boldsymbol{x}_0$



(b) Partial variable in $\boldsymbol{x}_2$

Figure 4: Partial variables in $\boldsymbol{x}_0$ and $\boldsymbol{x}_2$ of the practical system

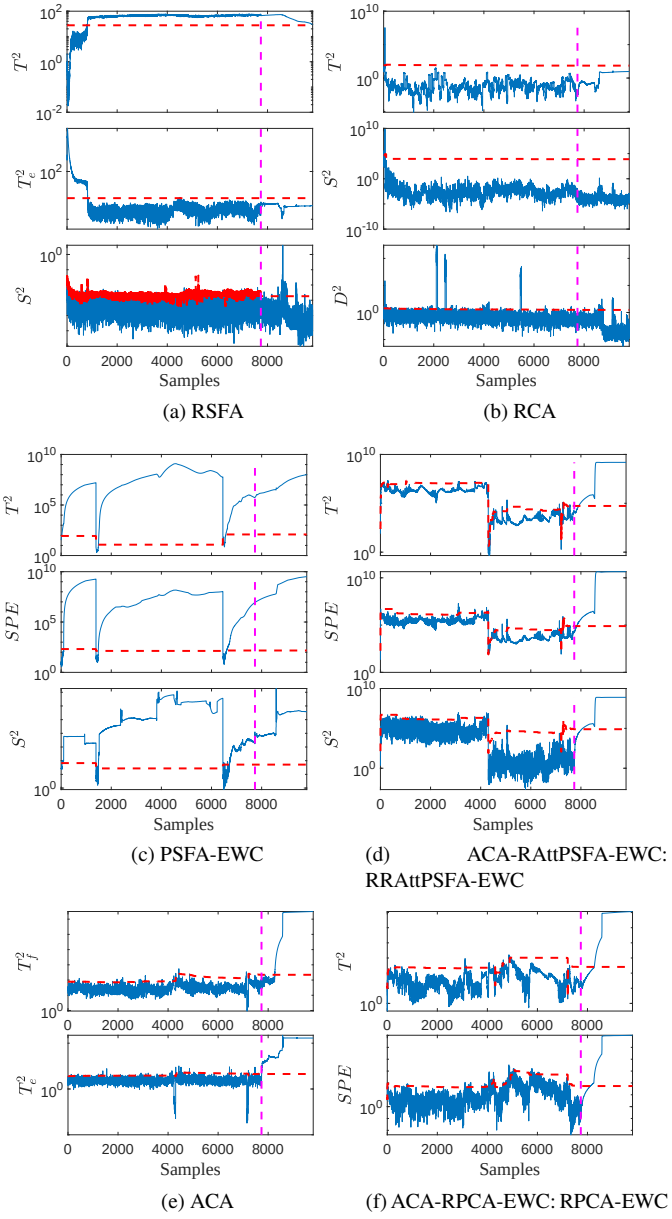
and vary with the real-time load, as depicted in Figure 3. Since it is hard to mix the coal evenly and the component of the same coal may vary with the time as environment, the long-term equilibrium may vary slowly for one mode. Therefore, this paper involved ACA to track the variation adaptively. Variables in $\boldsymbol{x}_0$ are stationary in each mode and may be regulated to another value when the coal changes, as described in Figure 4a. For instance, the outlet temperature is desired to remain the same for one coal. However, the actual value may change frequently around the set value owing to the controller and other factors. It is assigned to $\boldsymbol{x}_0$ based on the process mechanism and should be stationary for local modes. Variables in $\boldsymbol{x}_2$ may be affected by environment or change irregularly. For example, the inlet air temperature is closely linked with the environment and is nonstationary for local modes, as listed in Figure 4b.

For this practical coal pulverizing system, the partial variables are seriously affected by noise owing to the environmental disturbance and system noise, as depicted in Figure 3. The long-term equilibrium information from $\boldsymbol{x}_1$ is extracted after ACA and the remaining information with noise is represented as $\boldsymbol{x}_1\boldsymbol{W}_f^{\perp}$. As mentioned in Section 3.3, $\hat{\boldsymbol{x}}_2 = \left[\boldsymbol{x}_1\boldsymbol{W}_f^{\perp} \;\; \boldsymbol{x}_2\right]$ is processed by RAttPSFA and the contained noise is monitored by S^2 statistic.

4.2.3. Experimental results and discussion

One thousand normal samples are adopted initially for offline training. When a new normal mode arrives, 90 normal samples are used for offline training. Then the model is updated and used for online monitoring. During 30 minutes, the model of the last mode is utilized for online monitoring, which is utilized to illustrate the real-time monitoring performance.

The monitoring consequences of three cases are listed in Table 3. Owing to the limitation of paper length, the monitoring charts of Case 4 are provided as a representative, as described in Figure 5. Different from the other four methods,



(a) RSFA (b) RCA

(c) PSFA-EWC (d) ACA-RAttPSFA-EWC: RRAttPSFA-EWC

(e) ACA (f) ACA-RPCA-EWC: RPCA-EWC

Figure 5: Monitoring charts of Case 4

the mode information is required to be available for PSFA-EWC. RSFA is unable to monitor three cases accurately. The FARs of Cases 3 and 4 are higher than 85%, which indicate that RSFA fails to track the rapid and dramatical variations between modes and the normal changes may be misjudged as a fault. As shown in Figure 5a, the FAR of T^2 statistic is 89.55%. Conversely, the FDR of Case 5 is lower than 11%, which means that the fault is misidentified as normal variation. RCA can detect the fault in Case 5, in which the FDR is 98.18% and the FAR is lower than 3%. However, the FDRs of Cases 3 and 4 are lower than 6%. The normal nonstationary variations and the real fault could be separated by RCA (as shown in Figure 5b). PSFA-EWC also fails to monitor three cases and the FARs are

higher than 94%. As illustrated in Figure 5c, only a few data are collected to retrain the monitoring model when a new mode arrives, which is directly applied to online monitoring without updating. This model contains limited critical information of the operating mode, thus leading to terrible monitoring performance.

Since ACA is both used in ACA-RPCA-EWC and ACA-RAttPSFA-EWC, they share the same results of T_f^2 and T_e^2, as described in Figure 5e. The differences focus on the short-term dynamics, which are extracted by RPCA-EWC (Figure 5f) and RAttPSFA-EWC (Figure 5d). The mode can be identified by ACA and the FDRs of T_e^2 are higher than 96%. When a new mode arrives, 90 normal samples are used to establish initial CA, PCA-EWC and AttPSFA-EWC models, which would be updated respectively and recursively based on the forthcoming data. In other words, the ACA-RAttPSFA-EWC and ACA-RPCA-EWC monitoring models are corrected gradually for online applications. For Case 3, RPCA-EWC fails to detect the fault accurately and the FDRs are 0. Relatively speaking, RAttPSFA-EWC can detect the fault accurately, and the FDRs of T^2 and S^2 are higher than 90%. With regard to Case 4, the FDR of RAttPSFA-EWC is higher than 95%, while the FDRs of RPCA-EWC are lower than 82%. RAttPSFA-EWC and RPCA-EWC can provide excellent performance for Case 5. Although ACA-RPCA-EWC and ACA-RAttPSFA-EWC can identify the mode accurately due to the accurate ACA model, ACA-RAttPSFA-EWC can deliver more desirable performance than ACA-RPCA-EWC because nonlinear dynamic features are extracted deeply by RAttPSFA-EWC. Furthermore, uncertainly such as noise is considered and the proposed method enhances interpretability due to the probabilistic form.

The training and testing time are listed in Tables 4 and 5, which can reflect the computational complexity directly. Among four adaptive monitoring methods, the proposed ACA-RAttPSFA-EWC costs the most expensive computational resources. For Cases 3–5, the testing time for each sample is 0.3294, 0.7556 and 0.6923 second on average, respectively. The sampling interval is 20s and thus the online computational complexity is accepted for the proposed ACA-RAttPSFA-EWC method. For PSFA-EWC, the training time is far higher than that of other adaptive methods, even higher than the sum of training time and testing time. The online complexity of PSFA-EWC is the lowest since the parameters have already been estimated after the training procedure.

In conclusion, ACA-RAttPSFA-EWC provides the most excellent performance of the five methods, where the FDRs are satisfactory and the FARs are acceptable. Besides, the mode could be identified automatically without any human intervention for online applications, which makes it convenient for industrial systems.

5. Conclusion

This paper has introduced an intelligent adaptive monitoring method for multimode nonstationary processes, which can identify the mode automatically and account for measurement noise. The ACA algorithm extracts the long-term equilibrium features and the remaining dynamic information is further decomposed by the proposed RAttPSFA-EWC. The attention mechanism is adopted to focus on the global and local important information. AttPSFA-EWC has been proposed to handling the high-dimensional data for offline training procedure, which shares the similar framework with PSFA-EWC. Then, the parameters are updated recursively based on the forthcoming data for online monitoring. In comparison with several advanced methods using a numerical case and a practical coal pulverizing system, the effectiveness of ACA-RAttPSFA-EWC is validated.

Since regularization-based continual learning requires similarity among multiple modes and is suitable for short-term monitoring tasks, the continual learning ability of ACA-RAttPSFA-EWC would decrease if more diverse modes emerge continuously. Therefore, an adaptive monitoring method needs to be investigated to monitor long-term multiple nonstationary modes.

Acknowledgements

This work was supported by National Natural Science Foundation of China [grant numbers 62303114, 62033008, 62303090], the Fundamental Research Funds for the Central Universities, Natural Science Foundation of Jiangsu Province [grant number BK20230825], Natural Science Foundation of Sichuan Province [[grant numbers 2024NSFSC1480], the Postdoctoral Science Foundation of China [grant numbers 2023M740516], Zhishan Young Scholar of Southeast University.

References

[1] J. Li, K. Huang, D. Wu, Y. Liu, C. Yang, W. Gui. Hybrid variable dictionary learning for monitoring continuous and discrete variables in manufacturing processes. Control Eng Prac 149 (2024) 105970.

[2] J. Shang, M. Chen, H. Ji, D. Zhou, H. Zhang, M. Li. Dominant trend based logistic regression for fault diagnosis in nonstationary processes. Control Eng Prac 66 (2017) 156–168.

[3] H. Fan, C. Lu, X. Lai, S. Du, W. Yu, M. Wu. Adaptive monitoring for geological drilling process using neighborhood preserving embedding and Jensen–Shannon divergence. Control Eng Prac 134 (2023) 105476.

[4] X. Xu, J. Ding. Similarity and sparsity collaborative embedding and its application to robust process monitoring. Control Eng Prac 122 (2022) 105113.

[5] Q. Jiang, X. Yan. Multimode process monitoring using variational Bayesian inference and canonical correlation analysis. IEEE Trans Automat Sci Eng 16 (4) (2019) 1814–1824.

[6] Z. Chen, D. Zhou, E. Zio, T. Xia, E. Pan. Adaptive transfer learning for multimode process monitoring and unsupervised anomaly detection in steam turbines. Reliab Eng Syst Safe 234 (2023) 109162.

[7] M. Quiñones-Grueiro, A. Prieto-Moreno, C. Verde, O. Llanes-Santiago. Data-driven monitoring of multimode continuous processes: A review. Chemom Intell Lab Syst 189 (2019) 56–71.

[8] K. Zhang, K. Peng, S. Zhao, Z. Chen. A novel common and specific features extraction-based process monitoring approach with application to a hot rolling mill process. Control Eng Prac 104 (2020) 104628.

[9] J. Dong, C. Zhang, K. Peng. A new multimode process monitoring method based on a hierarchical Dirichlet process—hidden semi-Markov model with application to the hot steel strip mill process. Control Eng Prac 110 (2021) 104767.

[10] J. Zhang, D. Zhou, M. Chen, X. Hong. Continual learning for multimode dynamic process monitoring with applications to an ultra–supercritical thermal power plant. IEEE Trans Automat Sci Eng 20 (1) (2023) 137–150.

[11] W. Yu, C. Zhao, B. Huang. Recursive cointegration analytics for adaptive monitoring of nonstationary industrial processes with both static and dynamic variations. J Process Control 92 (2020) 319–332.

[12] C. Shang, F. Yang, B. Huang, D. Huang. Recursive slow feature analysis for adaptive monitoring of industrial processes. IEEE Trans Ind Electron 65 (11) (2018) 8895–8905.

[13] K. Huang, Y. Wu, C. Yang, G. Peng, W. Shen. Structure dictionary learning-based multimode process monitoring and its application to aluminum electrolysis process. IEEE Trans Automat Sci Eng 17(4) (2020) 1989–2003.

[14] W. Yu, C. Zhao. Recursive exponential slow feature analysis for fine-scale adaptive processes monitoring with comprehensive operation status identification. IEEE Trans Ind Informat 15 (6) (2019) 3311–3323.

[15] J. Chen, C. Zhao. Exponential stationary subspace analysis for stationary feature analytics and adaptive nonstationary process monitoring. IEEE Trans Ind Informat 17 (12) (2021) 8345–8356.

[16] J. Zhang, D. Zhou, M. Chen. Adaptive cointegration analysis and modified RPCA with continual learning ability for monitoring multimode nonstationary processes. IEEE Trans Cybern 53 (8) (2023) 4841–4854.

[17] J. Zhang, D. Zhou, M. Chen, X. Hong. Continual learning-based probabilistic slow feature analysis for monitoring multimode nonstationary processes. IEEE Trans Automat Sci Eng 21 (1) (2024) 733–745.

[18] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. IEEE Trans Pattern Anal Mach Intell 44 (7) (2022) 3366–3385.

[19] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, S. Wermter. Continual lifelong learning with neural networks: A review. Neural Netw 113 (2019) 54–71.

[20] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska. Overcoming catastrophic forgetting in neural networks. Proc Nat Acad Sci USA 114 (13) (2017) 3521–3526.

[21] G. M. van de Ven, H. T. Siegelmann, A. S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. Nature Commun 11 (1) (2020) 4069–4069.

[22] A. Mallya, S. Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7765–7773.

[23] F. Guo, C. Shang, B. Huang, K. Wang, F. Yang, D. Huang. Monitoring of operating point and process dynamics via probabilistic slow feature analysis. Chemom Intell Lab Syst 151 (151) (2016) 115–125.

[24] J. Zhang, M. Chen, X. Hong. Monitoring multimode nonlinear dynamic processes: An efficient sparse dynamic approach with continual learning ability. IEEE Trans Ind Informat 19 (7) (2023) 8029–8038.

[25] R. F. E. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55 (2) (1987) 251–276.

[26] S. Johansen. Statistical analysis of cointegration vectors. J Econ Dyn Control 12 (23) (1988) 231–254.

[27] L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, S. Nikitidis, M. Pantic. Probabilistic slow features for behavior analysis. IEEE Trans Neural Netw Learn Syst 27 (5) (2016) 1034–1048.

[28] S. Zhong, T. M. Khoshgoftaar, N. Seliya. Clustering-based network intrusion detection. Int J Reliab Qual Sa Eng 14 (02) (2007) 169–187.

[29] R. Turner, M. Sahani. A maximum-likelihood interpretation for slow feature analysis. Neural Comput 19 (4) (2007) 1022–1038.

[30] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39 (1) (1977) 1–22.

[31] O. Cappé, E. Moulines. On-line expectation–maximization algorithm for latent data models. J R Stat Soc B 71 (3) (2009) 593–613.