Multimodal continual learning for process monitoring: a novel weighted canonical correlation analysis with attention mechanism

Article

Accepted Version

Multimodal continual learning for process monitoring: a novel weighted canonical correlation analysis with attention mechanism

Jingxin Zhang, James Xiao, Maoyin Chen and Xia Hong, *Senior Member, IEEE*

***Abstract*—Aimed at sequential dynamic modes, a novel multimodal weighted canonical correlation analysis using attention mechanism (MWCCA-A) is introduced to derive a single model for process monitoring, by integrating two ideas of replay and regularization in continual learning. Under the assumption that data are received sequentially, subsets of data from past modes with dynamic features are selected and stored as replay data, which are utilized together with the current mode data for continual model parameter estimation. The weighted canonical correlation analysis is introduced to achieve appropriate weightings of past modes' replay data, so that the latent variables are extracted by maximizing the weighted correlation with its prediction via the attention mechanism. Specifically, replay data weightings are obtained via the probability density estimation from each mode. This is also beneficial in overcoming data imbalance amongst multiple modes and consolidating the significant features of past modes further. Alternatively, the proposed model also regularizes parameters based on its previous modes' importance, which is measured by synaptic intelligence. Meanwhile, the objective is decoupled into a regularization-related part and a replay-related part, to overcome the potentially unstable optimization trajectory of synaptic intelligence-based continual learning. In comparison with several multimode monitoring methods, the effectiveness of the proposed MWCCA-A approach is demonstrated by a continuous stirred tank heater, Tennessee Eastman process and a practical coal pulverizing system.**

***Index Terms*—Multimode dynamic process monitoring, weighted canonical correlation analysis, replay and regularization continual learning, synaptic intelligence, attention mechanism**

I. Introduction

Multimode process monitoring has become a pervasive challenge in industrial system modelling and has attracted wide interest [1]–[4]. Various complex modelling systems/processes generate a large amount of data, which are inherently dynamic and involve changes in system operating conditions. These may be related to the underlying system or process, such as raw materials, setting points, etc [5], [6]. Dynamic process monitoring has been researched in recent decades, including dynamic inner principal component analysis (DiPCA) [7]. DiPCA is robust to collinearity and delivers interpretability as well as prediction performance. However, DiPCA does not maximize correlation [8]. To overcome this issue, dynamic inner canonical correlation analysis (DiCCA) was presented to extract the most predictable information by maximizing the correlation between the variables and predictions [9].

Traditional multimode process methods generally divide multimode data into several clusters, followed by local monitoring models being constructed corresponding to each mode or a global model built for existing modes based on Bayesian inference [4]. Typically, a mixture of canonical variate analysis (MCVA) was presented for multimode dynamic processes, where the mode was identified by Gaussian mixture models and subsequently local dynamic models were built within each mode [1]. However, these methods assume that data from all potential modes are available. In practical industrial applications, sensing data are collected sequentially, and it is intractable to collect complete data to train a perfect model. When new modes arrive successively, traditional approaches need to store all historical data [1], [2], [10] and retrain the model from scratch without using the learned knowledge. It is therefore highly important to develop efficient methods to monitor sequential modes, with acceptable storage and computational resource requirements.

Recently, continual learning has been applied to multimode process monitoring and has achieved excellent performance [11]–[13], which assumes that data are generated successively, and the model is updated without forgetting the previously learned knowledge [14]–[17]. Three techniques have been researched to alleviate the catastrophic forgetting issue, including regularization, data replay and parameter isolation [18]. Parameter isolation approaches generally isolate parameters for specific tasks [19], and the capacity for new tasks is minimized to avoid saturation and ensure stable learning for future tasks. Regularization-based continual learning has already been applied to multimode process monitoring [11], [12], where an extra quadratic term is introduced to consolidate the previous knowledge to learn a new mode, via the estimation of the importance of model parameters [18]. Specifically, sparse principal component analysis (SPCA) with continual learning ability was proposed for multimode stationary processes [12], where the importance was measured by synaptic intelligence (SI) [20]. This method was denoted as SPCA–SI. Subsequently, a novel sparse DiPCA (SDiPCA)

This work was supported by National Natural Science Foundation of China [grant numbers 62303114, 62373213], Jiangsu Natural Science Foundation [grant number BK20230825], and the Fundamental Research Funds for the Central Universities.

Jingxin Zhang is with the Department of Automation, and also with Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China (e-mail: zjx18@tsinghua.org.cn).

James Xiao, Independent researcher, Vancouver, BC, Canada.

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.

was investigated for multimode dynamic processes [11], where modified SI (MSI) was presented to overcome the limitations of traditional SI. It is referred to as SDiPCA–MSI. However, it has been analyzed [21] that regularization-based continual learning using SI may incur unstable optimization trajectory due to improper hyperparameters. Meanwhile, regularization continual learning is limited to modes with high similarity and the performance may degrade abruptly when the modes are diverse [22]. Therefore, it may be inappropriate if applied to long-term monitoring tasks.

Alternatively, replay continual learning [14], [18], [22] is effective to address shortcomings of regularization continual learning and is suitable for long-term monitoring tasks. Data in raw format are stored or pseudo-samples may be generated from a generative model, and would be replayed when a new mode arrives. Multimode nonlinear SDiPCA (MNSDiPCA) was presented for sequential modes [13], where a small part of sensing data were selected and would be replayed together with the current data to retrain a new monitoring model. One potential issue is that the stored replay data or the current mode data play an equally important role in multimode processes. However, the statistics of selected replay data may differ from overall past mode data, or the data numbers amongst past modes and the current mode data may be very imbalanced; these factors will influence parameter estimates in the model and could reduce modelling performance.

Against this background, this work introduces a novel weighted canonical correlation analysis (WCCA) with attention mechanism (WCCA-A), which is applicable to a single dynamic mode initially and then extended to monitoring sequential dynamic modes. The relationship of dynamic latent variables in canonical correlation analysis [9] is characterized by attention mechanism. The proposed multimodal WCCA-A with continual learning ability is denoted as MWCCA-A. Several complementary strategies such as SI-based regularization are incorporated in MWCCA-A to alleviate catastrophic forgetting issues, in which the importance of model parameters is evaluated by SI [20] and tuned by [21]. Simultaneously, new replay approaches are proposed to enhance the continual learning ability subject to satisfying the constraints of computational resources. Specifically, some data from each mode are selected and stored when the training procedure finishes per mode. With each incoming new mode, sufficient current mode data and previously stored replay data are jointly utilized for retraining the model. To overcome the potential imbalance between replayed data and the current mode data, the weights of replayed data are estimated by probability density estimation and used within the MWCCA-A framework.

The contributions of this paper are outlined below:

a) This paper proposes a novel WCCA-A approach for multimode dynamic process monitoring (MWCCA-A), where replay and regularization continual learning techniques are adopted to consolidate the previous knowledge via sensing data and the learned model. It can achieve excellent long-term and short-term monitoring performance.
b) To mitigate data imbalance among multiple modes, the weightings of replay data are allocated by the Parzen window density estimation algorithm, which is calculated based on each mode respectively.
c) The objective function is divided into two parts to deal with the unstable optimization trajectory, which leads to less human intervention. Besides, the proposed MWCCA-A method can avoid the potentially unstable training procedure that is caused by SI-based regularization continual learning.

The remainder of this paper is organized below. Section II introduces the framework of the proposed approach. Section III introduces the procedure of proposed MWCCA-A, including replay data selection, weight allocation, attention key updating algorithm, optimization solver including estimation of parameter importance for regularization, and the monitoring procedure. In Section IV, several state-of-the-art multimode monitoring methods are discussed to highlight the virtues of the proposed method. The effectiveness of the proposed method is demonstrated by several industrial systems in Section V. The concluding remarks are given in Section VI.

II. A NOVEL FRAMEWORK OF MULTIMODE PROCESS MODEL COMBINING REGULARIZATION AND REPLAY

This section introduces a new framework of continual learning, combining regularization and replay, for multimode process model that is based on WCCA-A to extract dynamic features. The concept of multimode parameter regularization for continual learning is initially described, then the proposed WCCA-A model is investigated for dynamic processes, followed by a problem statement of the proposed MWCCA-A.

A. Multimode parameter regularization for continual learning

The regularization technique is shown to be effective for monitoring multiple sequential modes with continual learning capability [4], [11], [12]. Consider the task of monitoring multimode dynamic processes, with each of the modes being denoted as $\mathcal{M}_K$, for sequential modes $K = 1, 2, \ldots$. In contrast to the common approach of building local models for each mode followed by combining them as a global model, only a single adaptive model is obtained. While the model is updated with incoming data from being collected of $\mathcal{M}_K$, it can explain $\mathcal{M}_K$ as well as all previous modes.

Without loss of generality, the concept of regularization for continual learning [20] is initially introduced for an unspecified system model, based on the gradient descent algorithm in the context that the model parameters are updated between two consecutive modes per step. Denote the associated parameter vector as $\boldsymbol{\theta}$. When the Kth mode $\mathcal{M}_K$ arrives, normal data $\boldsymbol{X}_K^0$ are collected. Parameter regularization is designed using the concept of SI to control catastrophic forgetting [20], where a quadratic regularization term is added to the loss to penalize changes in important model parameters. Specifically, the regularized objective function of regularization continual learning at current mode $\mathcal{M}_K$ can be formulated as

$$J_{reg}^K = J_K(\boldsymbol{\theta}) + \frac{\lambda}{2} \sum_i \hat{\varpi}_i^{K-1} \left(\theta_i - \theta_{K-1,i}^*\right)^2 \qquad (1)$$

where $J_K(\boldsymbol{\theta})$ is a model specific objective function using new data information at the Kth mode. $\lambda > 0$ is a regularization hyperparameter, $\theta^*_{K-1,i}$ denotes the parameter estimate from the previous mode $\mathcal{M}_{K-1}$, and $\hat{\varpi}_i^{K-1}$ is an importance parameter from SI for continual learning, as detailed below.

Considering gradient descent method to solve (1), each element of $\boldsymbol{\theta}$ is updated by

$$\theta_i \leftarrow \theta_i - \eta \left(\nabla_{\theta_i} J_K + \lambda \hat{\varpi}_i^{K-1} \left(\theta_i - \theta^*_{K-1,i} \right) \right) \tag{2}$$

where $\eta > 0$ is a small learning rate. Reformulating (2), the parameter is updated by

$$\theta_i \leftarrow \underbrace{\left(1 - \eta\lambda\varpi_i^{K-1}\right)\theta_i + \left(\eta\lambda\varpi_i^{K-1}\right)\theta^*_{K-1,i}}_{\text{Interpolation current and previous values}} - \underbrace{\eta\frac{\partial J_K}{\partial \theta_i}}_{\text{mode derivative}} \tag{3}$$

The updating of θ contains the regularization-related part (the first term) and the optimization about the current mode (the second term) in (3). It can be observed from (3) that a new mode monitoring task is trained through moving model parameters along mode-specific derivatives, and the catastrophic forgetting issue is alleviated by adding an interpolation operation between current and previous model parameters [21]. The interpolation term minimizes the variation of important parameters to retain the previously learned knowledge.

To represent the influence of interpolation visually, we derive the total variation in mode parameters between the $(K-1)$th and Kth modes. According to (3), unrolling the parameter updates via recursive substitution for k iterations yields

$$\theta_i = \theta^*_{K-1,i} - \sum_{j=0}^{k-1} \underbrace{\left[\left(1 - \eta\lambda\varpi_i^{K-1}\right)^{k-j-1}\right]}_{\text{effective learning rate}} g_j \tag{4}$$

in which $g_j = \frac{\partial J_K}{\partial \theta_i}$ at jth iteration. However, it has already been analyzed in [21] that improper hyperparameter configurations often result in extrapolation of parameters and the training process may be unstable. To address this issue, the updating of θ is divided into two operations:

$$\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} J_K \tag{5a}$$

$$\theta_i \leftarrow (1 - r_i)\theta_i + r_i \theta^*_{K-1,i} \tag{5b}$$

where the relative importance r_i is defined as [21]

$$r_i = \frac{\sqrt{\hat{\varpi}_i^{K-1}}}{\sqrt{\hat{\varpi}_i} + \sqrt{\hat{\varpi}_i^{K-1}}} \tag{6}$$

and $\hat{\varpi}_i$ is calculated by SI as follows [20]

$$\boldsymbol{\varpi} = \sum_k \left(\left(-\nabla_{\boldsymbol{\theta}} J_K \left(\boldsymbol{\theta}_k \right) \right)^T \odot \left(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1} \right)^T \right)^T \tag{7}$$

where $\odot$ denotes the Khatri-Rao product. After the training procedure, each element of $\boldsymbol{\varpi}$ is normalized by [23]

$$\hat{\varpi}_i = \max\left(0, \frac{\varpi_i}{\left(\Delta\theta_i\right)^2 + \zeta}\right) \tag{8}$$

where $\Delta\theta_i$ is the total change of ith variable for mode $\mathcal{M}_K$. $\zeta > 0$ is added to avoid ill-conditioning issues and let $\zeta = 1e^{-8}$ in this paper. Once the optimization procedure finishes, $\hat{\boldsymbol{\varpi}}$ is determined. The importance measure is updated for the $(K+1)$th mode as follows:

$$\boldsymbol{\varpi}^K = \left(\boldsymbol{\varpi}^{K-1} + \hat{\boldsymbol{\varpi}}\right)/2 \tag{9}$$

Remark: We further explain the rationale to update parameters using (5) instead of (3). In most cases, a gradient descent method is adopted to optimize the objective (1) and the parameter is updated by (4). Here, $1 - \eta\lambda\varpi_i^{K-1}$ can be regarded as a learning rate [21]. Once an improper hyperparameter λ is selected, it may lead to $\eta\lambda\varpi_i^{K-1} > 1$. Then, the training procedure is unstable because $1 - \eta\lambda\varpi_i^{K-1} < 0$. To avoid this problem, the updating steps are divided into two parts, namely, mode-specific changes in (5a) and interpolation (5b). The parameter importance of previous modes is evaluated by r_i in (6) and the updating procedure is not affected by the hyperparameter λ, which also reduces the need for human intervention to select an appropriate hyperparameter λ.

B. WCCA-A model for a single mode dynamic process

The attention mechanism is an efficient technique to overcome catastrophic forgetting [24], [25]. It is also beneficial for extracting global and local important features and ignoring the unimportant information. In this section, a novel weighted CCA model with attention mechanism is proposed by constructing a set of dynamical latent attention variables, and then the dynamical relationship of attention variables is characterized by a vector autoregressive (VAR) model.

Let $\boldsymbol{X} = \{\boldsymbol{x}_k\}$, $k = 1, \ldots, N$ as time instance. N is the number of samples and $\boldsymbol{x} \in \Re^m$ is a sample. Consider an attention function $\mathcal{F}$: $\boldsymbol{x} \longmapsto \phi(\boldsymbol{x})$, and $\phi(\boldsymbol{x}) = \{\phi_i(\boldsymbol{x})\} \in \Re^M$ given by

$$\phi_i(\boldsymbol{x}) = -\frac{\|\boldsymbol{x} - \boldsymbol{c}_i\|^2}{d} \tag{10}$$

where $d > 0$ is a scaling hyperparameter and $\boldsymbol{C} = \{\boldsymbol{c}_i\}$, $i = 1, \ldots, M$ are a set of M keys.

Attention is the mapping [26]

$$\texttt{Attention}(\boldsymbol{x}, \boldsymbol{C}, \boldsymbol{w}) = \sum_{i=1}^{M} \texttt{softmax}(\boldsymbol{x}, \boldsymbol{C})_i w_i \tag{11}$$

in which

$$\texttt{softmax}(\boldsymbol{x}, \boldsymbol{C})_i = \frac{\exp(\phi_i(\boldsymbol{x}))}{\sum_{i=1}^{M} \exp(\phi_i(\boldsymbol{x}))} \tag{12}$$

For convenience, the function of $\texttt{softmax}(\cdot)$ is denoted as $\boldsymbol{x}_\phi$ and $\texttt{Attention}(\boldsymbol{x}, \boldsymbol{C}, \boldsymbol{w})$ is labelled by t. Through the attention mechanism (12), the mapped data are denoted as $\boldsymbol{X}_\phi \in \Re^{N \times M}$ and the kth sample is $\boldsymbol{x}_{\phi,k}$ correspondingly.

Define the latent attention variable at the kth instant as

$$t_k = \boldsymbol{x}_{\phi,k}^T \boldsymbol{w} \tag{13}$$

where $\boldsymbol{w} = [w_1, \ldots, w_M]^T$ is the weight vector with $\|\boldsymbol{w}\|_2 = 1$. Similar to DiCCA [8], the current latent attention variable

is represented by the past ones through a VAR model, namely,

$$t_k = \sum_{j=1}^{s} \beta_j t_{k-j} + r_k \tag{14}$$

where r_k is the Gaussian white noise at kth instant and s is the order of VAR model. According to (13) and (14), the prediction of dynamic latent attention variable is described by

$$\begin{aligned} \hat{t}_k &= \sum_{j=1}^{s} \boldsymbol{x}_{\phi,k-j}^T \boldsymbol{w} \beta_j \\ &= \left[\boldsymbol{x}_{\phi,k-1}^T \quad \cdots \quad \boldsymbol{x}_{\phi,k-s}^T\right] (\boldsymbol{\beta} \otimes \boldsymbol{w}) \end{aligned} \tag{15}$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_s]^T$ and $\|\boldsymbol{\beta}\|_2 = 1$. By extending the idea in DiCCA [8], [9], the proposed WCCA-A extracts dynamic features by maximizing the weighted correlation between the latent variable t_k and its prediction $\hat{t}_k$, namely,

$$\frac{\sum_{k=s+1}^{N} \omega_k t_k \hat{t}_k}{\sqrt{\sum_{k=s+1}^{N} \omega_k t_k^2} \sqrt{\sum_{k=s+1}^{N} \omega_k \hat{t}_k^2}} \tag{16}$$

where $\omega_k > 0$ are preset weights to each data instance. When all weights are equal, it is equivalent to unweighted correlation. For notational convenience, the description mentioned above is reformulated in a vector form. Define $\boldsymbol{X}_\phi^{(s+1)}$ and $\boldsymbol{Z}$ as

$$\boldsymbol{X}_\phi^{(j)} = \left[\boldsymbol{x}_{\phi,j} \;\; \boldsymbol{x}_{\phi,j+1} \;\; \cdots \;\; \boldsymbol{x}_{\phi,N-s+j-1}\right]^T, \;\; j = 1, \ldots, s+1 \tag{17}$$

$$\boldsymbol{Z} = \left[\boldsymbol{X}_\phi^{(s)} \;\; \boldsymbol{X}_\phi^{(s-1)} \;\; \cdots \;\; \boldsymbol{X}_\phi^{(1)}\right] \tag{18}$$

Then, the sub-vectors of latent variables are described as

$$\boldsymbol{t}_j = \boldsymbol{X}_\phi^{(j)} \boldsymbol{w} \in \Re^{N-s}, \;\; j = 1, \ldots, s+1 \tag{19}$$

The prediction $\hat{\boldsymbol{t}}_j$ of $\boldsymbol{t}_j$ is rewritten based on (15) into

$$\hat{\boldsymbol{t}}_{s+1} = \sum_{j=1}^{s} \beta_j \boldsymbol{t}_{s+1-j} = \sum_{j=1}^{s} \beta_j \boldsymbol{X}_\phi^{(s+1-j)} \boldsymbol{w} \tag{20}$$

According to the expression mentioned above, the objective (16) can be reformulated as

$$\max_{\boldsymbol{w},\boldsymbol{\beta}} \; \frac{\boldsymbol{t}_{s+1}^T \boldsymbol{\Omega} \hat{\boldsymbol{t}}_{s+1}}{\left\|\boldsymbol{\Omega}^{\frac{1}{2}} \boldsymbol{t}_{s+1}\right\| \left\|\boldsymbol{\Omega}^{\frac{1}{2}} \hat{\boldsymbol{t}}_{s+1}\right\|}$$

where $\boldsymbol{\Omega} = \mathtt{diag}\{\omega_{s+1}, \ldots, \omega_N\}$, which is equivalent to

$$\begin{aligned} &\max_{\boldsymbol{w},\boldsymbol{\beta}} \;\; \boldsymbol{t}_{s+1}^T \boldsymbol{\Omega} \hat{\boldsymbol{t}}_{s+1} \\ &s.t. \;\; \left\|\boldsymbol{\Omega}^{\frac{1}{2}} \boldsymbol{t}_{s+1}\right\|^2 = 1, \left\|\boldsymbol{\Omega}^{\frac{1}{2}} \hat{\boldsymbol{t}}_{s+1}\right\|^2 = 1 \end{aligned} \tag{21}$$

Let $\overline{\boldsymbol{X}}$ denote $\boldsymbol{X}_\phi^{(s+1)}$. By making use of (19) and (20), the objective of WCCA-A is designed as

$$\begin{aligned} &\min \;\; J(\boldsymbol{w},\boldsymbol{\beta}) = -\boldsymbol{w}^T \overline{\boldsymbol{X}}^T \boldsymbol{\Omega} \boldsymbol{Z} (\boldsymbol{\beta} \otimes \boldsymbol{w}) + \lambda_1 \boldsymbol{\beta}^T \boldsymbol{D} \boldsymbol{\beta} \\ &s.t. \;\; \left\|\boldsymbol{\Omega}^{\frac{1}{2}} \overline{\boldsymbol{X}} \boldsymbol{w}\right\|_2 = 1, \;\; \left\|\boldsymbol{\Omega}^{\frac{1}{2}} \boldsymbol{Z} (\boldsymbol{\beta} \otimes \boldsymbol{w})\right\|_2 = 1 \end{aligned} \tag{22}$$

where $\boldsymbol{D}$ is an unknown weighting matrix to make $\boldsymbol{\beta}$ sparse, and λ_1 is a hyperparameter and predefined by users. Sparse representation contributes to avoiding potential overfitting

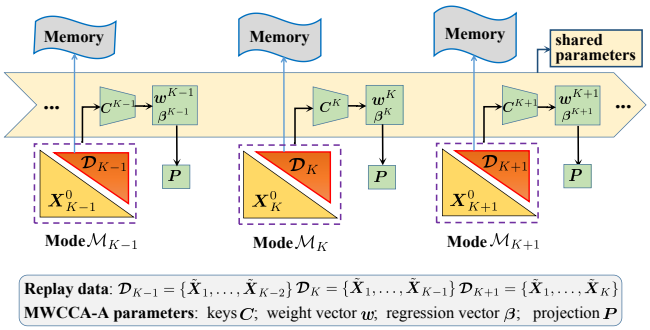

Fig. 1. Overview of MWCCA-A with regularization and replay continual learning

[11], [27] and alleviating catastrophic forgetting [22]. $\boldsymbol{\Omega}$ is an externally supplied constant matrix in above objective function, which can be set as an identity matrix in the case of single-mode data sets, which is equivalent to unweighted CCA with attention mechanism.

C. Problem Statement

This work aims to adaptively build a single MWCCA-A model, upon receiving sequential modes, by integrating both replay and regularization techniques. This presents a novel framework for a continual learning multimode monitoring approach in order to maintain the performance for all modes with acceptable storage and computing costs.

Considering that the regularized objective function (1) is integrated within the specific context of WCCA-A, we have $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{\beta}\}$ and the term $J_K(\boldsymbol{\theta})$ in (1) should be contributed by data set $\overline{\boldsymbol{X}}_K$ and $\boldsymbol{Z}_K$ from data $\boldsymbol{X}_K$ which follow (17) and (18), except that the subscript is added to indicate the mode index K. As illustrated in Fig. 1, in the proposed approach, $\boldsymbol{X}^K$ is formed jointly from data of the current mode K, and all previous stored replay data $\boldsymbol{\mathcal{D}}_K$ in memory, so that data in raw format of previous modes are replayed together with new data when a new mode arrives. Specifically as data from $\mathcal{M}_K$, $K > 1$ are received sequentially, the attention mechanism parameters $\{\boldsymbol{C}^K, \boldsymbol{w}^K\}$ and the VAR parameters $\boldsymbol{\beta}^K$ are shared parameters between neighborhood modes and updated from a prior model parameter set $\{\boldsymbol{C}^{K-1}, \boldsymbol{w}^{K-1}\}$ and $\boldsymbol{\beta}^{K-1}$, using data $\boldsymbol{X}_K^0$, as well as previous stored replay data $\boldsymbol{\mathcal{D}}_K$ from each previous mode. $\boldsymbol{w}^{K-1}$ and $\boldsymbol{\beta}^{K-1}$ are optimal parameters of the mode $\mathcal{M}_{K-1}$, that are obtained recursively as below. The projection matrix $\boldsymbol{P}$ is utilized to calculate the monitoring index in Section III.C.

According to (1), (22) and replay continual learning, the objective of MWCCA-A at Kth mode can be formulated as

$$\begin{aligned} &\min \;\; J_{total}^K(\boldsymbol{w},\boldsymbol{\beta}) = J^K(\boldsymbol{w},\boldsymbol{\beta}) + J_{reg}\left(\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{w}^{K-1},\boldsymbol{\beta}^{K-1}\right) \\ &s.t. \;\; \left\|\boldsymbol{\Omega}_K^{\frac{1}{2}} \overline{\boldsymbol{X}}^K \boldsymbol{w}\right\|_2 = 1, \;\; \left\|\boldsymbol{\Omega}_K^{\frac{1}{2}} \boldsymbol{Z}^K (\boldsymbol{\beta} \otimes \boldsymbol{w})\right\|_2 = 1 \end{aligned} \tag{23}$$

Similar to (17) and (18), construct $\overline{\boldsymbol{X}}^K$ and $\boldsymbol{Z}^K$ based on data $\boldsymbol{X}^K$ from the existing K modes. The replay-related term $J^K(\boldsymbol{w},\boldsymbol{\beta})$ is constructed by

$$J^K(\boldsymbol{w},\boldsymbol{\beta}) = -\boldsymbol{w}^T \left(\overline{\boldsymbol{X}}^K\right)^T \boldsymbol{\Omega}_K \boldsymbol{Z}^K (\boldsymbol{\beta} \otimes \boldsymbol{w}) + \lambda_1 \boldsymbol{\beta}^T \boldsymbol{D} \boldsymbol{\beta}$$

in which $\boldsymbol{\Omega}_K$ measures the importance of joint data $\boldsymbol{X}^K$. According to (1), the regularization-related term $J_{reg}\left(\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{w}^{K-1},\boldsymbol{\beta}^{K-1}\right)$ is formulated as

$$\begin{aligned} &J_{reg}\left(\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{w}^{K-1},\boldsymbol{\beta}^{K-1}\right) \\ =&\frac{\gamma_{1,K}}{2}\left(\boldsymbol{w}-\boldsymbol{w}^{K-1}\right)^T\boldsymbol{\Pi}_w\left(\boldsymbol{w}-\boldsymbol{w}^{K-1}\right) \\ &+\frac{\gamma_{2,K}}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{K-1}\right)^T\boldsymbol{\Pi}_\beta\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{K-1}\right) \end{aligned} \tag{24}$$

where $\boldsymbol{\varpi}_w^{K-1}$ and $\boldsymbol{\varpi}_\beta^{K-1}$ are the corresponding importance measures, and $\boldsymbol{\Pi}_w = diag(\boldsymbol{\varpi}_w^{K-1})$, $\boldsymbol{\Pi}_\beta = diag(\boldsymbol{\varpi}_\beta^{K-1})$. The parameters $\{\boldsymbol{w},\boldsymbol{\beta},\boldsymbol{\varpi}_w^K,\boldsymbol{\varpi}_\beta^K\}$ are updated by (5)–(9). $\gamma_{1,K}$ and $\gamma_{2,K}$ are regularization coefficients and predefined by users. Contrary to [4], [12], the training parameters are not affected by the hyperparameters through scaling down the regularization coefficient.

III. MWCCA-A FOR MULTIMODE PROCESS MONITORING

This section provides the details of MWCCA-A to monitor successive dynamic modes, integrating both replay and regularization continual learning for model estimation and its application for monitoring procedures. We start by outlining the technical ingredients that are: replay data selection, their weighting allocation and training data preparation. Then the unsupervised learning of keys $\boldsymbol{C}^K$ in the attention mechanism for each mode is introduced, followed by the parameter estimation of the proposed algorithm. Finally, the offline training and online monitoring phases are presented.

A. Subset replay data selection and weighting allocation

The MWCCA-A aims to build a single model for sequential modes with acceptable storage and computing costs. To reduce storage space and computation requirements, it is important to make the data selection algorithm as efficient as possible to store a selected fraction of the original data to be used as replay data. Alternatively once the replay data set is selected, it is desired that adequate weighting in MWCCA-A, as reflected $\boldsymbol{\Omega}_K$, can reflect the data distribution of each previous mode, for improved estimation. For example, in the case of an outlier being used as replay data, the weighting will reduce its impact on mode parameters.

1) Data selection: Define multiple modes as $\mathcal{M}_K$, $K = 1,2,\ldots$. To facilitate the exposition, assume that there are N_K normal samples in data matrix $\boldsymbol{X}_K^0 \in \Re^{N_K\times m}$ for each mode $\mathcal{M}_K$, which are normalized to zero means as well as unit variances, to yield $\boldsymbol{X}_K$. For each mode $\mathcal{M}_K$, some replay data are selected in order to present the operating condition without much redundancy. A novel data selection technique is presented by combining online k-means algorithm [28] and k nearest neighborhood (KNN). Specifically, k-means clustering is adopted to acquire several cluster centers, which can reflect the data distribution to a certain extent. Subsequently, KNN is used to find the corresponding samples in original data space that are nearest to the cluster centers. The procedure of data selection is summarized in Algorithm 1. Therefore, the selected data $\tilde{\boldsymbol{X}}_K$ can represent the original data, and between them there is sufficient dissimilarity. The number of clusters n remains the same for different modes and is generally set to be sufficiently large to cover the data space. It is possible that different clustering centers may share the same nearest samples. Thus, the number of selected data may be different when replay data are ensured to be different.

Algorithm 1 Data selection via online k-means and KNN

Input: Data $\boldsymbol{X}_K \in \Re^{N_K\times m}$, the number of clusters n, t_{total}.
Output: Representative data $\tilde{\boldsymbol{X}}_K$.

1: Initialize cluster centers $\boldsymbol{U} = \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n\}$, $iter = 1$, $\eta_1 = 0.05$, $n_1 = 2$.
2: For $iter = 1 : t_{total}$
For $k = 1 : N_K$
a) For each data $\boldsymbol{x}_k \in \boldsymbol{X}_K$, $i^* = \arg\min_i \quad \|\boldsymbol{x}_k - \boldsymbol{u}_i\|^2$, $i \in \{1,\ldots,n\}$;
b) Update the cluster center $\boldsymbol{u}_{i^*}$ as

$$\begin{aligned} \boldsymbol{u}_{i^*}^{new} &= \boldsymbol{u}_{i^*} - \eta_{iter}\frac{\partial L}{\partial \boldsymbol{u}_{i^*}} \\ &= \boldsymbol{u}_{i^*} + \eta_{iter}(\boldsymbol{x}_k - \boldsymbol{u}_{i^*}) \end{aligned}$$

c) Update the learning rate $\eta_{iter} = \eta_0(\frac{\eta_f}{\eta_0})^{\frac{iter}{t_{total}}}$, let $iter = iter + 1$;
3: The optimal cluster centers are denoted as $\boldsymbol{U}^K = \{\boldsymbol{u}_1^*,\ldots,\boldsymbol{u}_n^*\}$.
4: According to KNN, find n samples in $\boldsymbol{X}_K$ that are nearest to $\boldsymbol{u}_1^*,\ldots,\boldsymbol{u}_n^*$ correspondingly. Delete the redundant vectors and the rest vectors are denoted as $\tilde{\boldsymbol{X}}_K \in \Re^{n_K\times m}$.

2) Weight allocation: Obviously, data from previous modes and the current mode are generally imbalanced. Only a few representative samples from previous modes are selected and stored for replay continual learning, hence data size of previous modes could be much lower than that of the current mode. The significant features from previous modes may be forgotten because these features could be overwritten by the current mode, which motivates the need for data weighting. For each mode, a Parzen window density estimation algorithm [29] is utilized to generate a weighting for each replay data instance. For simplicity, denote $\boldsymbol{x}_k \in \Re^m$, $k = 1 : N_K$, as data samples of $\boldsymbol{X}_K$. The Parzen window probability density function (PDF) estimator [29] can be written as

$$\begin{aligned} PDF(\boldsymbol{x}) = &\frac{1}{N_K(2\pi)^{m/2}\prod\sigma_i} \\ &\times\sum_{k=1}^{N_K}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_k)\} \end{aligned} \tag{25}$$

where $\boldsymbol{\Sigma} = \text{diag}\{\sigma_1^2,...,\sigma_m^2\}$, σ_i is called the bandwidth and set using Scott's rule of thumb [30] that minimizes the mean integrated squared error to true, unknown density, as

$$\sigma_i \approx S_i N_K^{-1/(m+4)}$$

Algorithm 2 Replay data weighting using Parzen window PDF estimation

Input: Data $\boldsymbol{X}_K$, replay data set $\tilde{\boldsymbol{X}}_K = \{\tilde{\boldsymbol{x}}_1,\ldots,\tilde{\boldsymbol{x}}_{n_K}\}$.
Output: The corresponding weighting vector for replay data $\tilde{\boldsymbol{X}}_K$ as $\boldsymbol{q}_K = \{q(\tilde{\boldsymbol{x}}_i)\} \in \Re^{n_K}$.

1: Denote each data $\boldsymbol{x}_k \in \boldsymbol{X}_K$, $k = 1 : N_K$.
2: Find PDF for $\boldsymbol{X}_K$ using (25).
3: $q(\tilde{\boldsymbol{x}}_i) = \frac{N_K PDF(\tilde{\boldsymbol{x}}_i)}{\sum_{i=1}^{n_K} PDF(\tilde{\boldsymbol{x}}_i)}$
4: **return**

Algorithm 3 Updating $\boldsymbol{C}^K$ based on maximum likelihood

Input: Data $\boldsymbol{X}^K$, $\boldsymbol{C}^{K-1}$, $\eta = 0.1$, error ϵ.
Output: Key $\boldsymbol{C}^K = \{\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_M^*\}$.
1: Initialize $t = 1$, use $\boldsymbol{C}^{K-1}$ as the initial setting of $\boldsymbol{C}$, calculate the initial $L(0) = \sum_{k=1}^{N^K} L_k$ based on (26);
2: For each data $\boldsymbol{x}_k (k = 1, \ldots, N^K)$, update $\boldsymbol{c}_i$ using gradient ascent update of (27);
3: Calculate $L(t) = \sum_{k=1}^{N^K} L_k$, and L_k is calculated by (26);
4: Return to step 2 until $\|L(t) - L(t-1)\| < \epsilon$, let $t = t + 1$;
5: The optimal cluster centers are denoted as $\boldsymbol{C}^K = \{\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_M^*\}$.

in which S_i is the standard deviation of samples in the ith feature of mode data $\boldsymbol{X}_K$. The data weighting procedure is outlined in Algorithm 2.

3) Training data preparation and pre-training algorithm for $\boldsymbol{C}^K$: Assume that data from multiple modes are collected sequentially, replay data $\tilde{\boldsymbol{X}}_{K-1}$ have been selected after training the mode $\mathcal{M}_{K-1}$, together with their weighting $\boldsymbol{q}_{K-1}$ being obtained. Thus, $\boldsymbol{\mathcal{D}}_K = \left\{\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_{K-1}\right\}$ are available for previous modes. When Kth mode $\mathcal{M}_K$ arrives, normal data $\boldsymbol{X}_K^0$ are normalized as $\boldsymbol{X}_K$. Let $\boldsymbol{X}^K = \{\boldsymbol{\mathcal{D}}_K, \boldsymbol{X}_K\} \in \Re^{N^K \times m}$ be constructed, where N^K is the number of prepared training samples (current mode and replay modes) that are ready to be employed in training algorithms.

As summarized in Algorithm 3, the key $\boldsymbol{C}^K$ is updated via an unsupervised pre-training manner once a new mode appears, using instantaneous log-likelihood function (26)

$$L_k = \sum_{i=1}^{M} \log \texttt{softmax}(\boldsymbol{x}_k, \boldsymbol{C})_i \tag{26}$$

as

$$\begin{aligned} \boldsymbol{c}_i^{new} &= \boldsymbol{c}_i^{old} + \eta \frac{\partial}{\partial \boldsymbol{c}_i} L_k \\ \boldsymbol{c}_i^{new} &= \boldsymbol{c}_i^{new} / \sum_i \|\boldsymbol{c}_i^{new}\|, \end{aligned} \tag{27}$$

with

$$\delta \boldsymbol{c}_i = \frac{\partial}{\partial \boldsymbol{c}_i} L_k = 2(M \texttt{softmax}(\boldsymbol{x}, \boldsymbol{C})_i - 1) \frac{(\boldsymbol{x}_k - \boldsymbol{c}_i)}{d},$$

for all i if (10) is used. $\eta > 0$ is a small preset learning rate.

B. Parameter estimation of MWCCA-A algorithm

The proposed MWCCA-A method inherits the themes of replay and regularization continual learning, and thus it is appropriate for long-term and short-term monitoring tasks. The problem has been reformulated specifically in Section II-C.

Recall the objective function (23), we explain the construction of some critical matrices. Assume that data from multiple modes are collected and the model is trained in a sequential manner. When the mode $\mathcal{M}_K$ arrives, data of previous modes have been selected and replayed, followed by updating the key $\boldsymbol{C}^K$. Map data $\boldsymbol{X}^K$ to a high dimensional space by (11)–(12), and then calculate the mean $\boldsymbol{\mu}_K^\phi$ and variance $\boldsymbol{\Sigma}_K^\phi$. The preprocessed data are denoted as $\boldsymbol{X}_{\phi,K}$ with zero mean and unit variance. Similar to (17) and (18), construct $\boldsymbol{X}_{\phi,K}^{(j)}$ $(1 \le j \le s+1)$ and $\boldsymbol{Z}^K = \left[\boldsymbol{X}_{\phi,K}^{(s)} \;\; \boldsymbol{X}_{\phi,K}^{(s-1)} \;\; \cdots \;\; \boldsymbol{X}_{\phi,K}^{(1)}\right]$, and let $\overline{\boldsymbol{X}}^K$ denote $\boldsymbol{X}_{\phi,K}^{(s+1)}$. The weightings of replayed data have been estimated by Algorithm 2. $\boldsymbol{\Omega}_K$ is a diagonal weighting matrix with the dimension of an incremental data size along with that of $\overline{\boldsymbol{X}}^K$ (also $\boldsymbol{X}^K$) as K increases. It is defined as

$$\boldsymbol{\Omega}_K = \texttt{diag}\{\boldsymbol{q}_1^T, ..., \boldsymbol{q}_{K-1}^T, \alpha \boldsymbol{1}_{N_K}^T\} \tag{28}$$

where $\alpha > 0$ is a scaling parameter, balancing data importance between the current mode and past modes. It is predefined by users based on the expert experience and prior knowledge. $\boldsymbol{1}$ is an all-ones vector, reflecting equal weighting for data in the current mode.

Solution: The objective (23) is optimized by an augmented Lagrange multiplier method. To be consistent with (1), the augmented Lagrange function of (23) is divided into two parts

$$\tilde{J}_{total}^K(\boldsymbol{w}, \boldsymbol{\beta}) = \tilde{J}^K(\boldsymbol{w}, \boldsymbol{\beta}) + J_{reg}\left(\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{w}^{K-1}, \boldsymbol{\beta}^{K-1}\right) \tag{29}$$

where

$$\begin{aligned} \tilde{J}^K(\boldsymbol{w}, \boldsymbol{\beta}) =& J^K(\boldsymbol{w}, \boldsymbol{\beta}) + \rho_1 \left(\boldsymbol{w}^T \left(\overline{\boldsymbol{X}}^K\right)^T \boldsymbol{\Omega}_K \overline{\boldsymbol{X}}^K \boldsymbol{w} - 1\right)^2 \\ &+ \rho_2 \left((\boldsymbol{\beta} \otimes \boldsymbol{w})^T \left(\boldsymbol{Z}^K\right)^T \boldsymbol{\Omega}_K \boldsymbol{Z}^K (\boldsymbol{\beta} \otimes \boldsymbol{w}) - 1\right)^2 \end{aligned} \tag{30}$$

in which ρ_1 and ρ_2 are Lagrangian parameters, J_{reg} has been described in (24).

We aim to optimize (29) by a gradient descent method. Specifically, (5a) is realized by Nesterov-accelerated adaptive moment estimation (Nadam) [31] to optimize (30), which inherits the virtues of Adam and Nesterov accelerated gradient. Then, the regularization term (24) is updated by (5b).

For the proposed MWCCA-A, the gradients with regard to

Algorithm 4 Update parameters and importance at kth iteration: $[\boldsymbol{\theta}_{k+1}, \boldsymbol{\varpi}_{k+1}, \boldsymbol{m}_{k+1}, \boldsymbol{v}_{k+1}] = F\left(\boldsymbol{\theta}_k, \boldsymbol{m}_k, \boldsymbol{v}_k, \boldsymbol{g}_k, \boldsymbol{\theta}_0, \boldsymbol{\theta}_{pre}^*, \boldsymbol{\varpi}_k, \boldsymbol{\varpi}_{pre}\right)$

1: Update parameters by Nadam
   a) Update the first moment estimate: $\boldsymbol{m}_{k+1} = \mu_1 \boldsymbol{m}_k + (1 - \mu_1)\, \boldsymbol{g}_k$
   b) Update the second moment estimate: $\boldsymbol{v}_{k+1} = \mu_2 \boldsymbol{v}_k + (1 - \mu_2)\, \boldsymbol{g}_k \odot \boldsymbol{g}_k$
   c) Correct the first moment estimate: $\hat{\boldsymbol{m}}_{k+1} = \boldsymbol{m}_{k+1} / \left(1 - \mu_1^k\right)$
   d) Correct the second moment estimate: $\hat{\boldsymbol{v}}_{k+1} = \boldsymbol{v}_{k+1} / \left(1 - \mu_2^k\right)$
   e) Update each element of the parameter: $\hat{\theta}_{k+1,i} = \theta_{k,i} - \frac{\eta_2}{\sqrt{\hat{v}_{k+1,i} + \epsilon}} \left(\mu_1 \hat{m}_{k+1,i} + \frac{(1-\mu_1) g_{k,i}}{1 - \mu_1^k}\right)$, $i = 1, \cdots, n$, and $\hat{\boldsymbol{\theta}}_{k+1} = \left\{\hat{\theta}_{k+1,1}, \cdots, \hat{\theta}_{k+1,n}\right\}$
2: Correct parameters by considering the interpolation operation
   a) Update parameter importance: $\boldsymbol{\varpi}_{k+1} = \boldsymbol{\varpi}_k - \left(\boldsymbol{g}_k^T \odot \left(\hat{\boldsymbol{\theta}}_{k+1} - \boldsymbol{\theta}_k\right)^T\right)^T$
   b) Normalize importance: $\hat{\varpi}_{k+1,i} = \max\left(0, \frac{\varpi_{k+1,i}}{\left(\hat{\theta}_{k+1,i} - \theta_{0,i}\right)^2 + \zeta}\right)$
   c) Compute each element of relative importance: $r_i = \sqrt{\varpi_{pre,i}} / \left(\sqrt{\varpi_{pre,i}} + \sqrt{\hat{\varpi}_{k+1,i}}\right)$
   d) Correct parameters: $\theta_{k+1,i} = (1 - r_i)\hat{\theta}_{k+1,i} + r_i \theta_{pre,i}^*$
3: The updated parameter is $\boldsymbol{\theta}_{k+1} = \{\theta_{k+1,1}, \cdots, \theta_{k+1,n}\}$

Algorithm 5 Optimization procedure of MWCCA-A

Input: Data $\boldsymbol{X}_{\phi,K}$, parameters of mode $\mathcal{M}_{K-1}$ $\left\{\boldsymbol{W}_{\mathcal{M}_{K-1}}, \boldsymbol{\Gamma}_{\mathcal{M}_{K-1}}, \boldsymbol{\Pi}^{w}_{\mathcal{M}_{K-1}}, \boldsymbol{\Pi}^{\beta}_{\mathcal{M}_{K-1}}\right\}$

Output: Weight matrix $\boldsymbol{W}_{\mathcal{M}_K}$, regression matrix $\boldsymbol{\Gamma}_{\mathcal{M}_K}$, the importance measures $\boldsymbol{\Pi}^{w}_{\mathcal{M}_K}$ and $\boldsymbol{\Pi}^{\beta}_{\mathcal{M}_K}$, projection matrix $\boldsymbol{P}$, latent score matrix $\boldsymbol{T}$

for $j = 1, 2, \cdots, l$ **do**

1) Let $\boldsymbol{w}^{K-1}$, $\boldsymbol{\beta}^{K-1}$, $\hat{\boldsymbol{\varpi}}^{w}$, and $\hat{\boldsymbol{\varpi}}^{\beta}$ be the jth line of $\boldsymbol{W}_{\mathcal{M}_{K-1}}$, $\boldsymbol{\Gamma}_{\mathcal{M}_{K-1}}$, $\boldsymbol{\Pi}^{w}_{\mathcal{M}_{K-1}}$ and $\boldsymbol{\Pi}^{\beta}_{\mathcal{M}_{K-1}}$. $\boldsymbol{\Pi}_w = diag(\hat{\boldsymbol{\varpi}}^{w})$, $\boldsymbol{\Pi}_\beta = diag(\hat{\boldsymbol{\varpi}}^{\beta})$.
2) Random unit vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{w}_1$, $\boldsymbol{m}^{w}_1 = \boldsymbol{0}$, $\boldsymbol{v}^{w}_1 = \boldsymbol{0}$, $\boldsymbol{m}^{\beta}_1 = \boldsymbol{0}$, $\boldsymbol{v}^{\beta}_1 = \boldsymbol{0}$, $\boldsymbol{\varpi}^{w}_1 = \boldsymbol{0}$, $\boldsymbol{\varpi}^{\beta}_1 = \boldsymbol{0}$, $\boldsymbol{r}_w = \boldsymbol{1}$, $\boldsymbol{r}_\beta = \boldsymbol{1}$, $\rho_{1,1} = 0$, $\rho_{2,1} = 0$, $\mu_1 = 0.9$, $\mu_2 = 0.999$, $\alpha^{\rho} = 0.01$, $\eta = 0.001$, $\epsilon = 10^{-8}$, $\mu_1 = 0.9$, $\mu_2 = 0.999$, $\eta_2 = 0.001$, $k = 1$;
3) Construct $\boldsymbol{X}^{(j)}_{\phi,K}$ $(1 \le j \le s+1)$ and $\boldsymbol{Z}_K$ in accordance with (17) and (18);
4) Calculate optimal $\boldsymbol{w}$ and $\boldsymbol{\beta}$ based on the correlation of prediction:
while the objective (29) is not converged **do**
   a) Update parameters about $\boldsymbol{w}$, $[\boldsymbol{w}_{k+1}, \boldsymbol{\varpi}^{w}_{k+1}, \boldsymbol{m}^{w}_{k+1}, \boldsymbol{v}^{w}_{k+1}] = F\left(\boldsymbol{w}_k, \boldsymbol{m}^{w}_k, \boldsymbol{v}^{w}_k, \nabla_{\boldsymbol{w}}\tilde{J}^{K}\left(\boldsymbol{w}_k, \boldsymbol{\beta}_k\right), \boldsymbol{w}_1, \boldsymbol{w}^{K-1}, \boldsymbol{\varpi}^{w}_k, \hat{\boldsymbol{\varpi}}_w\right)$ in **Algorithm 4**, and $\nabla_{\boldsymbol{w}}\tilde{J}^{K}\left(\boldsymbol{w}_k, \boldsymbol{\beta}_k\right)$ is calculated by (31);
   b) Update parameters about $\boldsymbol{\beta}$, $\left[\boldsymbol{\beta}_{k+1}, \boldsymbol{\varpi}^{\beta}_{k+1}, \boldsymbol{m}^{\beta}_{k+1}, \boldsymbol{v}^{\beta}_{k+1}\right] = F\left(\boldsymbol{\beta}_k, \boldsymbol{m}^{\beta}_k, \boldsymbol{v}^{\beta}_k, \nabla_{\boldsymbol{\beta}}\tilde{J}^{K}\left(\boldsymbol{w}_{k+1}, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_1, \boldsymbol{\beta}^{K-1}, \boldsymbol{\varpi}^{\beta}_k, \hat{\boldsymbol{\varpi}}_\beta\right)$ in **Algorithm 4**, and $\nabla_{\boldsymbol{\beta}}\tilde{J}^{K}\left(\boldsymbol{w}_{k+1}, \boldsymbol{\beta}_k\right)$ is calculated by (32);
   c) Update the Lagrangian parameters: $\rho_{1,k+1} = \rho_{1,k} + \alpha^{\rho}\nabla_{\rho_1}\tilde{J}^{K}\left(\boldsymbol{w}_{k+1}, \boldsymbol{\beta}_{k+1}\right)$, $\rho_{2,k+1} = \rho_{2,k} + \alpha^{\rho}\nabla_{\rho_2}\tilde{J}^{K}\left(\boldsymbol{w}_{k+1}, \boldsymbol{\beta}_{k+1}\right)$;
   d) Update the weighted matrix $\boldsymbol{D}$ by (33) and let $k = k + 1$;
end while
5) The weighted vector is $\boldsymbol{w}^{*}_j$, the regression coefficient is $\boldsymbol{\beta}^{*}_j$. Let $\boldsymbol{t}_j = \boldsymbol{X}_{\phi,K}\boldsymbol{w}^{*}_j$, $\boldsymbol{p}_j = \boldsymbol{X}^{T}_{\phi,K}\boldsymbol{t}_j / \boldsymbol{t}^{T}_j\boldsymbol{t}_j$, deflate $\boldsymbol{X}_{\phi,K}$ as $\boldsymbol{X}_{\phi,K} := \boldsymbol{X}_{\phi,K} - \boldsymbol{t}_j\boldsymbol{p}^{T}_j$;
6) The importance measures are normalized by (8), and denoted as $\hat{\boldsymbol{\varpi}}^{w}_j$ and $\hat{\boldsymbol{\varpi}}^{\beta}_j$.

end for

$\boldsymbol{W}_{\mathcal{M}_K} = [\boldsymbol{w}^{*}_1\ \boldsymbol{w}^{*}_2\ \cdots\ \boldsymbol{w}^{*}_l]$, $\boldsymbol{\Gamma}_{\mathcal{M}_K} = [\boldsymbol{\beta}^{*}_1\ \boldsymbol{\beta}^{*}_2\ \cdots\ \boldsymbol{\beta}^{*}_l]$, $\boldsymbol{\Pi}^{w}_{\mathcal{M}_K} = [\hat{\boldsymbol{\varpi}}^{w}_1\ \hat{\boldsymbol{\varpi}}^{w}_2\ \cdots\ \hat{\boldsymbol{\varpi}}^{w}_l]$, $\boldsymbol{\Pi}^{\beta}_{\mathcal{M}_K} = \left[\hat{\boldsymbol{\varpi}}^{\beta}_1\ \hat{\boldsymbol{\varpi}}^{\beta}_2\ \cdots\ \hat{\boldsymbol{\varpi}}^{\beta}_l\right]$, $\boldsymbol{P} = [\boldsymbol{p}_1\ \boldsymbol{p}_2\ \cdots\ \boldsymbol{p}_l]$, $\boldsymbol{T} = [\boldsymbol{t}_1\ \boldsymbol{t}_2\ \cdots\ \boldsymbol{t}_l]$

each parameter are calculated by

$$\begin{aligned}&\nabla_{\boldsymbol{w}}\tilde{J}^{K}\left(\boldsymbol{w}, \boldsymbol{\beta}\right)\\ =&-\left(\boldsymbol{G}_{K,\boldsymbol{\beta}} + \boldsymbol{G}^{T}_{K,\boldsymbol{\beta}}\right)\boldsymbol{w} + 4\rho_1\left(\boldsymbol{w}^{T}\boldsymbol{G}_{K,x}\boldsymbol{w} - 1\right)\boldsymbol{G}_{K,x}\boldsymbol{w}\\ &+ 4\rho_2\left(\boldsymbol{w}^{T}\boldsymbol{G}_{K,z\beta}\boldsymbol{w} - 1\right)\boldsymbol{G}_{K,z\beta}\boldsymbol{w}\end{aligned} \tag{31}$$

$$\begin{aligned}\nabla_{\boldsymbol{\beta}}\tilde{J}^{K}\left(\boldsymbol{w}, \boldsymbol{\beta}\right) =&-\left(\boldsymbol{I}_s \otimes \boldsymbol{w}\right)^{T}\left(\boldsymbol{Z}^{K}\right)^{T}\boldsymbol{\Omega}_K\overline{\boldsymbol{X}}^{K}\boldsymbol{w} + 2\lambda_1\boldsymbol{D}\boldsymbol{\beta}\\ &+ 4\rho_2\left(\boldsymbol{\beta}^{T}\boldsymbol{G}_{K,zw}\boldsymbol{\beta} - 1\right)\boldsymbol{G}_{K,zw}\boldsymbol{\beta}\end{aligned} \tag{32}$$

where $\boldsymbol{G}_{K,z\beta} = \left(\boldsymbol{\beta}\otimes\boldsymbol{I}_M\right)^{T}\left(\boldsymbol{Z}^{K}\right)^{T}\boldsymbol{\Omega}_K\boldsymbol{Z}^{K}\left(\boldsymbol{\beta}\otimes\boldsymbol{I}_M\right)$, $\boldsymbol{G}_{K,zw} = \left(\boldsymbol{I}_s \otimes \boldsymbol{w}\right)^{T}\left(\boldsymbol{Z}^{K}\right)^{T}\boldsymbol{\Omega}_K\boldsymbol{Z}^{K}\left(\boldsymbol{I}_s \otimes \boldsymbol{w}\right)$, $\boldsymbol{G}_{K,x} = \left(\overline{\boldsymbol{X}}^{K}\right)^{T}\boldsymbol{\Omega}_K\overline{\boldsymbol{X}}^{K}$, $\boldsymbol{G}_{K,\beta} = \left(\overline{\boldsymbol{X}}^{K}\right)^{T}\boldsymbol{\Omega}_K\boldsymbol{Z}^{K}\left(\boldsymbol{\beta}\otimes\boldsymbol{I}_M\right)$. $\boldsymbol{I}_s$ and $\boldsymbol{I}_M$ are identity matrices with s and M dimensions, respectively.

The updating procedure of $\boldsymbol{w}$ and $\boldsymbol{\beta}$ is summarized in Algorithm 4. ρ_1 and ρ_2 are updated to accelerate the convergence rate. The gradients are calculated by

$$\nabla_{\rho_1}\tilde{J}^{K}\left(\boldsymbol{w}, \boldsymbol{\beta}\right) = \left(\boldsymbol{w}^{T}\left(\overline{\boldsymbol{X}}^{K}\right)^{T}\boldsymbol{\Omega}_K\overline{\boldsymbol{X}}^{K}\boldsymbol{w} - 1\right)^{2}$$

$$\nabla_{\rho_2}\tilde{J}^{K}\left(\boldsymbol{w}, \boldsymbol{\beta}\right) = \left(\left(\boldsymbol{\beta}\otimes\boldsymbol{w}\right)^{T}\left(\boldsymbol{Z}^{K}\right)^{T}\boldsymbol{\Omega}_K\boldsymbol{Z}^{K}\left(\boldsymbol{\beta}\otimes\boldsymbol{w}\right) - 1\right)^{2}$$

The detailed solution of MWCCA-A is outlined in Algorithm 5. During the optimization procedure, the matrix $\boldsymbol{D}$ is updated to avoid potential overfitting as follows [27]

$$\begin{aligned}\boldsymbol{D}^{k+1} =& diag\left\{d^{k+1}_1, d^{k+1}_2, \cdots, d^{k+1}_s\right\}\\ d_{k+1,j} =& \frac{1}{|\beta_{k+1,j}| + \epsilon}, \quad j = 1, \cdots, s\end{aligned} \tag{33}$$

where $\beta_{k+1,j}$ is the jth element of $\boldsymbol{\beta}_{k+1}$, $\epsilon = 1e^{-6}$ is added to avoid the potential ill-conditioning issue.

Importance measure: Recall (7), the importance measures for parameters $\boldsymbol{w}$ and $\boldsymbol{\beta}$ are calculated by

$$\boldsymbol{\varpi}^{w} = \sum_k\left(\left(-\nabla_{\boldsymbol{w}}\tilde{J}^{K}\left(\boldsymbol{w}_{k+1}, \boldsymbol{\beta}_{k+1}\right)\right)^{T} \odot \left(\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\right)^{T}\right)^{T}$$

$$\boldsymbol{\varpi}^{\beta} = \sum_k\left(\left(-\nabla_{\boldsymbol{\beta}}\tilde{J}^{K}\left(\boldsymbol{w}_{k+1}, \boldsymbol{\beta}_{k+1}\right)\right)^{T} \odot \left(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\right)^{T}\right)^{T}$$

After finishing the training procedure, $\boldsymbol{\varpi}^{w}$ and $\boldsymbol{\varpi}^{\beta}$ are normalized by (8), and denoted as $\hat{\boldsymbol{\varpi}}^{w}$ and $\hat{\boldsymbol{\varpi}}^{\beta}$ respectively. Then, the importance measures are updated and ready for the $(K+1)$th mode:

$$\boldsymbol{\varpi}^{K}_{w} = \left(\boldsymbol{\varpi}^{K-1}_{w} + \hat{\boldsymbol{\varpi}}_w\right)/2 \tag{34}$$

$$\boldsymbol{\varpi}^{K}_{\beta} = \left(\boldsymbol{\varpi}^{K-1}_{\beta} + \hat{\boldsymbol{\varpi}}_\beta\right)/2 \tag{35}$$

Remark: This work unifies replay and regularization-based continual learning, and it is appropriate for long-term and short-term monitoring tasks. A small amount of original data are selected based on the concept of nearest neighborhoods and would be replayed when a new mode arrives. Data nearest to the clustering centers can represent the most efficient information, and the data distribution is utilized to measure the

weight. Compared with data selection in [13], this work also considers data imbalance among multiple modes so that it is robust to outliers which are not used in replay. Simultaneously, MWCCA-A slows down learning for the parameters that are significant to previous modes, and the importance is evaluated by synaptic intelligence. Contrary to [12], the optimization procedure is stable and is not affected by the hyperparameters $\gamma_{1,K}$ and $\gamma_{2,K}$, as analyzed in Section II-A. The importance between the current mode and all previous modes is measured by hyperparameter α, which should be determined by prior knowledge and expert experience.

We briefly compare with parameter isolation methods [19] and explain that it is easier to adopt replay and regularization continual learning into the CCA framework. For online applications, it is necessary to judge whether the mode appears before. This work preserves common features by regularization continual learning and extracts specific features from diverse modes via replay continual learning. For parameter-isolation methods, part of parameters and resources are allocated for each specific monitoring mode and the mode needs to be identified for online monitoring, which may be challenging and presents an open problem for future research.

C. MWCCA-A for process monitoring

Similar to [7], [11], define $\boldsymbol{R} = \boldsymbol{W}_{\mathcal{M}_K}\left(\boldsymbol{P}^T\boldsymbol{W}_{\mathcal{M}_K}\right)^{-1}$, $\boldsymbol{T} = \boldsymbol{X}_{\phi,K}\boldsymbol{R}$. Form $\boldsymbol{T}_i$, $i = 1, \cdots, s+1$, from $\boldsymbol{T}$ in the same way with (17). Similar to (14), we establish the relationship between the latent score $\boldsymbol{T}_{s+1}$ and past $\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_s$ [7]

$$\begin{aligned}\boldsymbol{T}_{s+1} =& \sum_{i=1}^{s} \boldsymbol{T}_i \boldsymbol{\Phi}_{s+1-i} + \boldsymbol{E} \\ =& \bar{\boldsymbol{T}}_s \boldsymbol{\Phi} + \boldsymbol{E}\end{aligned}$$

Algorithm 6 Offline training procedure of MWCCA-A

Input: Data $\boldsymbol{X}_K$ and $\boldsymbol{X}^K$, weights $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{K-1}\}$, keys $\boldsymbol{C}^{K-1}$, weight matrix $\boldsymbol{W}_{\mathcal{M}_{K-1}}$, regression matrix $\boldsymbol{\Gamma}_{\mathcal{M}_{K-1}}$, importance measure $\boldsymbol{\Pi}^w_{\mathcal{M}_{K-1}}$ and $\boldsymbol{\Pi}^\beta_{\mathcal{M}_{K-1}}$

Output: Representative data $\tilde{\boldsymbol{X}}_K$, weight $\boldsymbol{q}_K$, keys $\boldsymbol{C}^K$, mean $\boldsymbol{\mu}^\phi_K$ and variance $\boldsymbol{\Sigma}^\phi_K$, $\boldsymbol{W}_{\mathcal{M}_K}$, $\boldsymbol{\Gamma}_{\mathcal{M}_K}$, $\boldsymbol{\Pi}^w_{\mathcal{M}_K}$, $\boldsymbol{\Pi}^\beta_{\mathcal{M}_K}$, $\boldsymbol{P}$, thresholds J_{th,T^2_φ} and J_{th,T^2_c}.

1: According to **Algorithm 3**, update $\boldsymbol{C}^K$ based on $\boldsymbol{X}^K$ and $\boldsymbol{C}^{K-1}$;
2: Map data $\boldsymbol{X}^K$ to a high-dimensional space via (10) and (12), which are denoted as $\boldsymbol{X}^0_{\phi,K}$;
3: Calculate the mean $\boldsymbol{\mu}^\phi_K$ and variance $\boldsymbol{\Sigma}^\phi_K$ of $\boldsymbol{X}^0_{\phi,K}$. Normalize data $\boldsymbol{X}^0_{\phi,K}$ to zero mean and unit variance, which are denoted as $\boldsymbol{X}_{\phi,K}$;
4: Construct the weight matrix $\boldsymbol{\Omega}_K$ via (28), calculate the optimal parameters $\boldsymbol{W}_{\mathcal{M}_K}$, $\boldsymbol{\Gamma}_{\mathcal{M}_K}$, $\boldsymbol{\Pi}^w_{\mathcal{M}_K}$, $\boldsymbol{\Pi}^\beta_{\mathcal{M}_K}$, $\boldsymbol{P}$ and $\boldsymbol{T}$ via Algorithm 5;
5: Calculate two statistics by (38) and (41), and the thresholds J_{th,T^2_φ} and J_{th,T^2_c} are estimated by KDE;
6: Select representative data $\tilde{\boldsymbol{X}}_K$ via online k-means and KNN in Algorithm 1;
7: Allocate the weight $\boldsymbol{q}_K$ of replay data $\tilde{\boldsymbol{X}}_K$ via Algorithm 2.

Algorithm 7 Online monitoring procedure of MWCCA-A

1: $\boldsymbol{x}^0$ is preprocessed by its mean and variance, and denoted as $\boldsymbol{x}$;
2: Map $\boldsymbol{x}$ to a high dimensional space via (10) and (12), the mapped $\boldsymbol{x}^0_\phi$ is preprocessed by $\boldsymbol{\mu}^\phi_K$ and variance $\boldsymbol{\Sigma}^\phi_K$, and denoted as $\boldsymbol{x}_\phi$;
3: Calculate the latent variables $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_s$ through $\boldsymbol{t} = \boldsymbol{x}_\phi \boldsymbol{R}$, and construct $\bar{\boldsymbol{t}}_s = [\boldsymbol{t}_1 \ \ \boldsymbol{t}_2 \ \cdots \ \boldsymbol{t}_s]$;
4: Predict the latent attention variable $\hat{\boldsymbol{t}}_{s+1}$ by (36), and calculate the dynamic prediction error by (37) and the static residual by (40);
5: Compute test statistics based on (38) and (41);
6: Evaluate the operating state: two statistics are lower than thresholds, the process is normal; otherwise, faulty.

where $\bar{\boldsymbol{T}}_s = [\boldsymbol{T}_1 \ \ \boldsymbol{T}_2 \ \cdots \ \boldsymbol{T}_s]$, $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_s \ \ \boldsymbol{\Phi}_{s-1} \ \cdots \ \boldsymbol{\Phi}_1]$. The least squares estimate for $\boldsymbol{\Phi}$ is

$$\hat{\boldsymbol{\Phi}} = \left(\bar{\boldsymbol{T}}_s^T \bar{\boldsymbol{T}}_s\right)^{-1} \bar{\boldsymbol{T}}_s^T \boldsymbol{T}_{s+1}$$

After $\hat{\boldsymbol{\Phi}}$ is obtained, $\boldsymbol{T}_{s+1}$ is predicted by

$$\hat{\boldsymbol{T}}_{s+1} = \bar{\boldsymbol{T}}_s \hat{\boldsymbol{\Phi}} \tag{36}$$

The dynamic residual matrix $\boldsymbol{V}$ is computed by:

$$\boldsymbol{V} = \boldsymbol{T} - \hat{\boldsymbol{T}}_{s+1} \tag{37}$$

Since the latent score matrix $\boldsymbol{T}$ may be dynamic, monitoring it directly would lead to high false alarm rate. Therefore, a monitoring index is built through $\boldsymbol{V}$ and defined as

$$T^2_\varphi = \boldsymbol{v}^T \boldsymbol{\Phi}_v \boldsymbol{v} \tag{38}$$

$$\boldsymbol{\Phi}_v = \frac{\boldsymbol{P}_v \boldsymbol{\Lambda}_v^{-1} \boldsymbol{P}_v^T}{J_{th,T^2_v}} + \frac{\boldsymbol{I} - \boldsymbol{P}_v \boldsymbol{P}_v^T}{J_{th,\mathrm{SPE_v}}} \tag{39}$$

where $\boldsymbol{P}_v$ is the principal component matrix by principal component analysis (PCA), and $\boldsymbol{\Lambda}_v = \frac{1}{N_K - s - 1}\boldsymbol{V}^T\boldsymbol{V}$. J_{th,T^2_v} and $J_{th,\mathrm{SPE_v}}$ are the thresholds of two statistics T^2_v and $\mathrm{SPE_v}$ based on PCA, respectively. T^2_v and SPE_v are the monitoring statistics and calculated based on PCA, where the projection matrix is $\boldsymbol{P}_v$ and eigenvalues are contained in $\boldsymbol{\Lambda}_v$. The static prediction error is

$$\boldsymbol{E} = \boldsymbol{X}_{s+1} - \boldsymbol{T}_{s+1} \boldsymbol{P}^T \tag{40}$$

Similar to (38), an index is designed to monitor the static error

$$T^2_c = \boldsymbol{e}^T \boldsymbol{\Phi}_c \boldsymbol{e} \tag{41}$$

$$\boldsymbol{\Phi}_c = \frac{\boldsymbol{P}_r \boldsymbol{\Lambda}_r^{-1} \boldsymbol{P}_r^T}{J_{th,T^2_r}} + \frac{\boldsymbol{I} - \boldsymbol{P}_r \boldsymbol{P}_r^T}{J_{th,\mathrm{SPE}_r}} \tag{42}$$

where $\boldsymbol{P}_r$ is the principal component matrix. Perform PCA on $\boldsymbol{E}$, then $\boldsymbol{E} = \boldsymbol{T}_r \boldsymbol{P}_r^T + \boldsymbol{E}_r$ and $\boldsymbol{\Lambda}_r = \frac{1}{N_K - s - 1}\boldsymbol{T}_r^T \boldsymbol{T}_r$. J_{th,T^2_r} and J_{th,SPE_r} are the thresholds of T^2_r and SPE_r.

Thresholds are estimated by kernel density estimation (KDE) [11], which are denoted as J_{th,T^2_φ} and J_{th,T^2_c} respectively. The offline training and online monitoring procedures are summarized in Algorithms 6 and 7, respectively. Fault detection rate (FDR), false alarm rate (FAR) and detection delay (DD) are utilized to evaluate the monitoring performance. FDR

and FAR are calculated as follows:

$$\mathrm{FDR} = \frac{\text{number of samples } (J > J_{th} | f \neq 0)}{\text{total samples } (f \neq 0)} \times 100\% \quad (43)$$

$$\mathrm{FAR} = \frac{\text{number of samples } (J > J_{th} | f = 0)}{\text{total samples } (f = 0)} \times 100\% \quad (44)$$

where J indicates the monitoring statistic and J_{th} is the corresponding threshold. If the process operates normally, $f = 0$; otherwise, $f \neq 0$. DD refers to the number of samples that the fault is detected later than the practical faulty time.

IV. Discussion and comparative experiment design

A. Comparison and discussion

The proposed MWCCA-A is compared with SDiPCA–MSI [11], SPCA–SI [12], MNSDiPCA [13], MCVA [1] and finite Gaussian mixture model (FGMM) [32]. These methods furnish continual learning ability except MCVA and FGMM, where regularization or replay technique is adopted to overcome the catastrophic forgetting issue of a single model for multiple modes. For MCVA and FGMM, the process data are assumed to be from different clusters and each cluster corresponds to a mode [32], which would be characterized by a Gaussian component. Bayesian inference is then adopted to derive an integrated global monitoring consequence for multimode processes. MCVA built local CVA monitoring models within each cluster and local statistics are calculated instead of local probability index. To be consistent with the other methods, two local monitoring statistics based on PCA are constructed within each Gaussian component for FGMM, instead of Mahalanobis distance. MCVA and FGMM need to store complete data of existing modes and retrain the model from scratch when a new mode arrives.

We focus on association and distinctions of four continual learning-based methods. They construct a single model for sequential modes, where significant features of new modes are extracted while consolidating the information from previous modes. We discuss these methods from five aspects:

a) *The manner of preserving information from previous modes.* SPCA–SI and SDiPCA–MSI adopt regularization-based continual learning, where the learned knowledge is consolidated by slowing down the learning rate of certain parameters when the model is updated. SPCA–SI utilized traditional SI to measure the importance of SPCA model parameters. The importance is easily influenced by initial setting of optimization issue. Aimed at this limitation, modified SI was proposed to estimate the importance of SDiPCA model parameters. However, it has been analyzed in Section II-A that improper hyperparameter settings would cause the unstable training procedure. Replay continual learning methods store representative data from previous modes and extract significant features from data in raw format, which would be replayed when a new mode arrives. MNSDiPCA only adopts the intention of replay continual learning. Regularization and replay continual learning are adopted simultaneously in MWCCA-A, where parameter importance is measured by traditional SI and the parameters are updated by (5) to avoid potential negative influence of improper hyperparameter configuration.

b) *Data selection and storage.* SPCA–SI and SDiPCA–MSI only use the learned knowledge from previous modes and the current mode data, which would be discarded once the training procedure finishes. Thus, there is no need to select and store representative data. MNSDiPCA utilized cosine similarity to reduce data redundancy, thus outliers may be selected and influence the extraction of dynamic features. MWCCA-A selects replay data based on online k-means and KNN, and the weights of these data are evaluated by Parzen window PDF, which can characterize the data distribution. Since MNSDiPCA and MWCCA-A store a fraction of data from previous modes, they cost a little more storage resources than SPCA–SI and SDiPCA–MSI.

c) *Data preprocessing.* For replay continual learning, replay data and the current mode data are integrated as a new data matrix. MNSDiPCA projected the reconstructed data onto a high dimensional space via a polynomial function to cope with nonlinearity, and the dimension of the mapped data is fixed. MWCCA-A maps the reconstructed data to a high dimensional space through attention mechanism, where the similarity is measured by negative Euclidean distance and the keys are estimated by maximum likelihood estimation. Then, replay data are allocated to different weights as mentioned in (28) before extracting multimode features. Contrary to MNSDiPCA, the dimension of mapped data is flexible and determined by prior knowledge.

d) *Applications.* SPCA–SI and SDiPCA–MSI require data similarity among different modes and are suitable for short-term monitoring tasks. MNSDiPCA can monitor multiple diverse modes via a single model and be applied to long-term monitoring tasks. The MWCCA-A model is retrained based on raw data and the learned knowledge from all existing modes, and thus it inherits the virtues of both continual learning techniques and can provide excellent performance for long-term and short-term monitoring tasks. In addition, SPCA–SI was investigated for multimode stationary processes, while the rest methods are presented for multimode dynamic processes.

e) *Online computational complexity.* The online computational complexity is measured by the term *flam* to present operation counts in Table I, which contains one addition and one multiplication [11], [33]. Similar to the proposed MWCCA-A, m is the dimension of original data, s is the auto-regressive order and l is the number of latent variables. For each sampling instance, SPCA–SI costs the least computational resource. Generally, let $M > m$ to characterize the nonlinear relationship, thus the computational complexity of MWCCA-A is higher than that of SDiPCA–MSI. When $M > \frac{m^2+3m}{2}$, the computational complexity of MWCCA-A is higher than that of MNSDiPCA. Note that the online computational complexity of these four methods is irrelevant to the number of existing modes K. However, the complexity of MCVA and FGMM would increase with the successive emergence of modes in future.

TABLE I
COMPARISON OF ONLINE COMPUTATIONAL COMPLEXITY

Algorithm	Complexity (*flam*)
MWCCA-A	$M^2 + (2m + 2l + 2)M + (s + 1)l^2 + 3l + 3m$
SDiPCA–MSI	$m^2 + (l + 3)m + (s + 1)l^2 + 3l$
SPCA–SI	$2m^2 + 3m$
MNSDiPCA	$\frac{m^4}{4} + \frac{5m^3}{2} + (\frac{25}{4} + l)m^2 + (3l + 6)m + (s + 1)l^2 + 3l$
MCVA	$8(s^2m^2 + sm)K + 2K$
FGMM	$(m^2 + 4m)K + 2K$

B. Comparative experiment design

In this paper, four dynamic modes arrive sequentially and the simulation scheme is summarized in Table II. SDiPCA–MSI [11], SPCA–SI [12], MNSDiPCA [13], MCVA [1] and FGMM [32] are adopted as comparisons to illustrate the superiority of the proposed MWCCA-A method. FDR (%), FAR (%) and DD are considered to evaluate the performance. It is not useful to consider detection delay when the FAR is higher than 20%. Note that the dash in the "Training data" column indicates the same data as the row above.

Similar to [11]–[13], Situations 1–19 are designed to illustrate the continual learning ability of MWCCA-A and the catastrophic forgetting issue of WCCA-A in multimode dynamic processes. Take the first two modes as an instance to explain the experiment. When the second mode $\mathcal{M}_2$ arrives, data $\boldsymbol{X}_2$, replay data $\boldsymbol{\mathcal{D}}_2$ ($\tilde{\boldsymbol{X}}_1$) of the previous mode and the previously learned knowledge $\mathcal{A}$ are utilized to train the model $\mathcal{C}$ and extract the multimode dynamic features. It is desired that the monitoring model $\mathcal{C}$ can monitor two modes $\mathcal{M}_1$ and $\mathcal{M}_2$ simultaneously, as the Situations 2 and 3 stated. Situation 4 is designed to illustrate the effectiveness of WCCA-A for a single dynamic mode. Situation 5 is utilized to illustrate the catastrophic forgetting of WCCA-A for multimode processes, namely, the monitoring model trained for one mode $\mathcal{M}_2$ fails to monitor another mode $\mathcal{M}_1$. When the modes $\mathcal{M}_3$ and $\mathcal{M}_4$ arrive one after another, similar to Situations 2–5, Situations 6–11 and Situations 12–19 are designed to evaluate the performance of MWCCA-A and WCCA-A.

SDiPCA–MSI, SPCA–SI and MNSDiPCA are adopted to compare the continual learning ability of MWCCA-A. SDiPCA–MSI and SPCA–SI were presented on the basis of regularization continual learning, where parameters important to previous modes are expected to change insignificantly to preserve the previously learned knowledge, as Situations 20–29 and Situations 30–39 designed. MNSDiPCA filters representative data based on cosine similarity, which would be replayed together with current mode data for retraining. Situations 40–49 are designed to reflect the continual learning ability of MNSDiPCA. MCVA is a typical multimode process monitoring approach, where the local monitoring models are built and then a global model is constructed based on Bayesian fusion. As Situations 50–67 shown, the MCVA model and FGMM model are retrained from scratch based on complete historical data when a new mode appears. They need expensive storage resources with the successive emergence of dynamic modes.

V. EXPERIMENTS AND SIMULATION ANALYSIS

This paper utilizes five methods in Section IV-A to compare with the proposed MWCCA-A. Continuous stirred tank heater (CSTH), Tennessee Eastman process (TEP) and a practical coal pulverizing system are adopted to illustrate the effectiveness of MWCCA-A.

A. CSTH

The CSTH process is a popular benchmark for multimode dynamic process monitoring, which aims to mix hot water and cold water [5], [11]. Level, temperature and flow are controlled by PI controllers and six interdependent variables are adopted for monitoring. Detailed description could refer to [34]. This paper considers the abnormality from temperature and the settings are summarized in Table III, where data are collected in a sequential manner. For each mode, 1000 normal samples are collected, and 1000 testing samples are generated including 500 normal samples and 500 faulty samples. The fault amplitude is 0.1.

The monitoring results of Case 1 are summarized in Table II. The proposed MWCCA-A method can deliver distinguished performance, where the FDRs are 100% and the FARs are lower than 0.70%. The FARs of Situations 1 and 3 are 3.0% and 0.40%, which indicates that the learned knowledge of mode $\mathcal{M}_2$ is beneficial to monitoring the previous mode $\mathcal{M}_1$. This phenomenon represents backward transfer learning ability, namely, the information of the future mode $\mathcal{M}_2$ is valuable for monitoring the previous mode $\mathcal{M}_1$. The FARs of Situations 2 and 16 are 0.40% and 3.40%, which reflects the forward transfer learning ability of MWCCA-A, namely, the features of previous modes are valuable to enhance the performance of the future mode $\mathcal{M}_4$. WCCA-A fails to monitor multiple modes, where the FARs of Situations 10, 17–19 are higher than 13.50%. The catastrophic forgetting issue is reflected in the monitoring model when one mode cannot detect the fault in another mode accurately. SDiPCA–MSI delivers outstanding performance except Situations 24 and 27. SPCA–SI is unable to monitor this multimode process, where the FDRs are lower than 80%. Similarly to SDiPCA–MSI, SPCA–SI can offer excellent performance except Situation 44. MNSDiPCA fails to deliver excellent detection accuracy, because the FARs of Situations 44 and 47 are not less than 10%. MCVA cannot monitor this process, for which the FDRs are lower than 96% and the FARs are higher than 11%. FGMM can monitor this case accurately, where the FDRs are 100% and the FARs are not higher than 3.0%.

MWCCA-A and MNSDiPCA need to select and store partial sets of data from previous modes, and the amount of replay data is listed in Table IV. Note that part of samples may be selected through Algorithm 1 more than once, duplicate data would be eliminated and thus the number of final representative data is different for four modes. MNSDiPCA selects fewer than 10% of normal samples based on cosine similarity. For MWCCA-A, fewer than 4% of samples are filtered from original training samples, which can further reduce the storage cost. According to Parzen window probability density estimation, different weights are allocated to replay

TABLE II
SIMULATION SCHEMES AND MONITORING RESULTS FOR CSTH, TEP AND THE COAL PULVERIZING SYSTEM

	Methods	Training data	Testing mode	Model label	CSTH			TEP						Coal pulverizing system		
					Case 1			Case 2			Case 3			Case 4		
					FDR	FAR	DD	FDR	FAR	DD	FDR	FAR	DD	FDR	FAR	DD
Situation 1	WCCA-A	$\boldsymbol{X}_1$	$\mathcal{M}_1$	$\mathcal{A}$	100	3.00	0	99.01	0.50	15	99.60	0.50	4	100	33.73	-
Situation 2	MWCCA-A	$\boldsymbol{X}_2$, $\boldsymbol{\mathcal{D}}_2$+ $\mathcal{A}$	$\mathcal{M}_2$	$\mathcal{B}$	100	0.40	0	97.36	0.50	20	99.93	0.50	0	100	0.26	0
Situation 3	MWCCA-A	-	$\mathcal{M}_1$	$\mathcal{B}$	100	0.40	0	99.08	2.75	14	99.67	2.75	0	100	0.83	0
Situation 4	WCCA-A	$\boldsymbol{X}_2$	$\mathcal{M}_2$	$\mathcal{C}$	100	1.80	0	97.82	0.75	20	100	0.75	0	100	0.66	0
Situation 5	WCCA-A	-	$\mathcal{M}_1$	$\mathcal{C}$	100	5.80	0	99.08	4.25	14	99.67	4.25	4	100	14.08	0
Situation 6	MWCCA-A	$\boldsymbol{X}_3$, $\boldsymbol{\mathcal{D}}_3$ + $\mathcal{B}$	$\mathcal{M}_3$	$\mathcal{E}$	100	0.40	0	98.81	0.25	16	99.93	0.25	1	100	0	0
Situation 7	MWCCA-A	-	$\mathcal{M}_1$	$\mathcal{E}$	100	0.40	0	98.81	0.25	16	98.22	3.50	0	100	0.59	0
Situation 8	MWCCA-A	-	$\mathcal{M}_2$	$\mathcal{E}$	100	0.40	0	97.03	6.75	20	99.87	6.75	0	100	0.52	0
Situation 9	WCCA-A	$\boldsymbol{X}_3$	$\mathcal{M}_3$	$\mathcal{F}$	100	1.60	0	98.95	0.25	16	100	0.25	0	100	0.31	0
Situation 10	WCCA-A	-	$\mathcal{M}_1$	$\mathcal{F}$	100	13.60	0	99.01	20.75	-	98.75	16.25	0	100	5.33	0
Situation 11	WCCA-A	-	$\mathcal{M}_2$	$\mathcal{F}$	100	3.80	0	97.96	36.25	-	99.21	28.75	-	100	66.97	-
Situation 12	MWCCA-A	$\boldsymbol{X}_4$, $\boldsymbol{\mathcal{D}}_4$ +$\mathcal{E}$	$\mathcal{M}_4$	$\mathcal{F}$	100	0.40	0	98.95	0.25	6	99.67	0.25	0	100	5.71	0
Situation 13	MWCCA-A	-	$\mathcal{M}_1$	$\mathcal{G}$	100	0.40	0	99.01	3.25	15	99.67	3.25	4	100	0.47	0
Situation 14	MWCCA-A	-	$\mathcal{M}_2$	$\mathcal{G}$	100	0.60	0	97.03	2.00	22	99.93	2.25	0	100	0.26	0
Situation 15	MWCCA-A	-	$\mathcal{M}_3$	$\mathcal{G}$	100	0.40	0	99.08	5.50	8	100	5.50	0	100	0	0
Situation 16	WCCA-A	$\boldsymbol{X}_4$	$\mathcal{M}_4$	$\mathcal{H}$	100	3.40	0	99.14	0.50	6	99.74	0.50	0	100	25.52	0
Situation 17	WCCA-A	-	$\mathcal{M}_1$	$\mathcal{H}$	100	14.40	0	99.14	5.50	13	99.67	5.50	4	100	0.36	0
Situation 18	WCCA-A	-	$\mathcal{M}_2$	$\mathcal{H}$	100	37.80	-	97.89	2.75	20	100	2.75	0	100	0.26	0
Situation 19	WCCA-A	-	$\mathcal{M}_3$	$\mathcal{H}$	100	35.20	-	99.14	15.75	8	100	16.00	0	100	0.20	0
Situation 20	SDiPCA	$\boldsymbol{X}_1$	$\mathcal{M}_1$	$\mathcal{I}$	100	4.20	0	98. 48	0.75	14	98.95	0.75	12	100	48.76	-
Situation 21	SDiPCA–MSI	$\boldsymbol{X}_2$ + $\mathcal{I}$	$\mathcal{M}_2$	$\mathcal{J}$	100	2.00	0	97.43	0.75	20	99.47	1.00	0	100	0.26	0
Situation 22	SDiPCA–MSI	-	$\mathcal{M}_1$	$\mathcal{J}$	100	5.20	0	99.08	2.75	14	99.41	2.75	1	100	50.89	-
Situation 23	SDiPCA–MSI	$\boldsymbol{X}_3$ + $\mathcal{J}$	$\mathcal{M}_3$	$\mathcal{K}$	100	0.60	0	97.76	0.75	16	99.54	0.75	3	100	5.51	0
Situation 24	SDiPCA–MSI	-	$\mathcal{M}_1$	$\mathcal{K}$	100	10.60	0	98.48	18.50	12	99.41	18.50	3	100	0.71	0
Situation 25	SDiPCA–MSI	-	$\mathcal{M}_2$	$\mathcal{K}$	100	4.00	0	96.84	30.25	-	99.74	30.50	-	100	1.57	0
Situation 26	SDiPCA–MSI	$\boldsymbol{X}_4$ + $\mathcal{K}$	$\mathcal{M}_4$	$\mathcal{L}$	100	1.40	0	97.50	0.75	7	99.67	0.75	0	100	24.33	-
Situation 27	SDiPCA–MSI	-	$\mathcal{M}_1$	$\mathcal{L}$	100	14.40	0	99.08	6.75	14	99.87	6.75	1	100	0.47	0
Situation 28	SDiPCA–MSI	-	$\mathcal{M}_2$	$\mathcal{L}$	100	5.60	0	95.52	3.00	21	99.93	3.00	0	100	0.26	0
Situation 29	SDiPCA–MSI	-	$\mathcal{M}_3$	$\mathcal{L}$	100	2.00	0	98.88	25.00	-	100	25.00	-	100	0	0
Situation 30	SPCA	$\boldsymbol{X}_1$	$\mathcal{M}_1$	$\mathcal{N}$	67.60	3.40	0	98.88	2.50	17	99.08	2.50	10	100	27.22	-
Situation 31	SPCA–SI	$\boldsymbol{X}_2$ + $\mathcal{N}$	$\mathcal{M}_2$	$\mathcal{O}$	59.80	4.40	0	95.99	2.75	23	99.74	2.75	3	99.68	0	0
Situation 32	SPCA–SI	-	$\mathcal{M}_1$	$\mathcal{O}$	70.00	3.60	0	98.82	3.50	18	99.01	3.50	10	100	0.12	0
Situation 33	SPCA–SI	$\boldsymbol{X}_3$ + $\mathcal{O}$	$\mathcal{M}_3$	$\mathcal{P}$	51.00	4.60	0	98.68	1.25	20	99.74	1.25	4	99.34	0	3
Situation 34	SPCA–SI	-	$\mathcal{M}_1$	$\mathcal{P}$	74.40	3.60	0	98.82	3.00	18	99.01	3.00	10	100	21.30	-
Situation 35	SPCA–SI	-	$\mathcal{M}_2$	$\mathcal{P}$	63.60	4.80	0	95.66	3.25	23	99.67	3.25	3	99.68	1.18	1
Situation 36	SPCA–SI	$\boldsymbol{X}_4$ + $\mathcal{P}$	$\mathcal{M}_4$	$\mathcal{Q}$	54.20	5.00	0	97.89	1.25	7	98.42	1.25	4	100	21.97	-
Situation 37	SPCA–SI	-	$\mathcal{M}_1$	$\mathcal{Q}$	79.60	4.60	0	98.82	2.50	18	98.75	2.50	12	100	0	0
Situation 38	SPCA–SI	-	$\mathcal{M}_2$	$\mathcal{Q}$	67.80	6.20	0	94.47	1.25	23	99.47	1.25	6	99.68	0	1
Situation 39	SPCA–SI	-	$\mathcal{M}_3$	$\mathcal{Q}$	58.40	6.00	0	98.68	0.25	20	99.67	0.25	5	99.34	0	3
Situation 40	NSDiPCA	$\bar{\boldsymbol{X}}_1$	$\mathcal{M}_1$	$\mathcal{R}$	100	8.40	0	99.08	1.75	14	99.74	1.75	4	100	60.71	-
Situation 41	MNSDiPCA	$\boldsymbol{X}_2$, $\bar{\boldsymbol{X}}_1$	$\mathcal{M}_2$	$\mathcal{S}$	99.40	0.60	0	98.02	1.50	20	100	1.75	0	100	0.39	0
Situation 42	MNSDiPCA	-	$\mathcal{M}_1$	$\mathcal{S}$	100	8.80	0	99.14	2.25	4	99.74	2.25	4	100	39.17	-
Situation 43	MNSDiPCA	$\boldsymbol{X}_3$, $\bar{\boldsymbol{X}}_2$, $\bar{\bar{\boldsymbol{X}}}_1$	$\mathcal{M}_3$	$\mathcal{T}$	96.18	0.60	0	99.01	1.50	15	99.93	1.50	1	100	0	0
Situation 44	MNSDiPCA	-	$\mathcal{M}_1$	$\mathcal{T}$	100	12.20	0	99.08	1.50	14	99.67	1.50	4	100	33.25	-
Situation 45	MNSDiPCA	-	$\mathcal{M}_2$	$\mathcal{T}$	100	6.80	0	98.02	1.75	20	100	1.75	0	100	30.93	-
Situation 46	MNSDiPCA	$\boldsymbol{X}_4$, $\bar{\boldsymbol{X}}_3$, $\bar{\bar{\boldsymbol{X}}}_2$, $\bar{\bar{\bar{\boldsymbol{X}}}}_1$	$\mathcal{M}_4$	$\mathcal{U}$	97.38	0.60	0	99.41	0.75	4	99.60	0.75	0	100	52.91	-
Situation 47	MNSDiPCA	-	$\mathcal{M}_1$	$\mathcal{U}$	100	10.00	0	99.08	1.25	14	99.60	1.25	4	100	33.02	-
Situation 48	MNSDiPCA	-	$\mathcal{M}_2$	$\mathcal{U}$	100	3.20	0	97.82	1.75	20	99.87	1.75	2	100	0.66	0
Situation 49	MNSDiPCA	-	$\mathcal{M}_3$	$\mathcal{U}$	100	1.20	0	99.01	2.25	15	99.93	2.25	1	100	0	0
Situation 50	MCVA	$\boldsymbol{X}_1, \boldsymbol{X}_2$	$\mathcal{M}_1$	$\mathcal{V}$	92.51	18.00	0	96.24	1.25	13	95.64	1.25	13	100	36.80	-
Situation 51	MCVA	-	$\mathcal{M}_2$	$\mathcal{V}$	95.95	44.00	-	83.43	1.25	28	97.62	1.25	9	100	4.33	0
Situation 52	MCVA	$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	$\mathcal{M}_1$	$\mathcal{W}$	85.83	13.80	0	96.04	1.00	13	95.71	1.00	13	100	35.38	-
Situation 53	MCVA	-	$\mathcal{M}_2$	$\mathcal{W}$	91.70	37.60	-	81.85	1.00	30	97.03	1.00	9	100	4.98	0
Situation 54	MCVA	-	$\mathcal{M}_3$	$\mathcal{W}$	91.09	42.40	-	79.41	0	15	99.01	0	3	100	0.31	0
Situation 55	MCVA	$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	$\mathcal{M}_1$	$\mathcal{X}$	80.36	11.20	0	95.38	0.25	26	96.04	0.25	23	100	34.67	-
Situation 56	MCVA	-	$\mathcal{M}_2$	$\mathcal{X}$	88.66	34.20	-	73.66	0	0	96.24	0	11	100	2.23	0
Situation 57	MCVA	-	$\mathcal{M}_3$	$\mathcal{X}$	87.25	39.20	-	83.10	0.25	15	98.94	0.25	5	100	65.01	-
Situation 58	MCVA	-	$\mathcal{M}_4$	$\mathcal{X}$	89.27	19.80	0	90.03	1.00	9	92.94	1.00	4	100	62.76	-
Situation 59	FGMM	$\boldsymbol{X}_1, \boldsymbol{X}_2$	$\mathcal{M}_1$	$\mathcal{Y}$	100	2.40	0	98.88	0	17	99.41	0	7	100	46.98	-
Situation 60	FGMM	-	$\mathcal{M}_2$	$\mathcal{Y}$	100	0.80	0	95.99	4.25	3	99.74	4.25	3	99.68	0	1
Situation 61	FGMM	$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	$\mathcal{M}_1$	$\mathcal{Z}$	100	2.60	0	98.95	1.25	16	99.67	1.25	4	100	70.89	-
Situation 62	FGMM	-	$\mathcal{M}_2$	$\mathcal{Z}$	100	0.80	0	5.72	0	8	2.30	0	67	99.68	3.41	1
Situation 63	FGMM	-	$\mathcal{M}_3$	$\mathcal{Z}$	100	0.80	0	98.75	3.00	19	99.61	3.00	6	99.78	31.13	-
Situation 64	FGMM	$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	$\mathcal{M}_1$	Ξ	100	3.00	0	98.88	1.50	17	99.54	1.50	7	100	58.70	-
Situation 65	FGMM	-	$\mathcal{M}_2$	Ξ	100	0.80	0	76.84	0	29	94.08	0	9	99.68	0.39	1
Situation 66	FGMM	-	$\mathcal{M}_3$	Ξ	100	0.80	0	98.75	3.50	19	99.61	3.50	6	100	25.64	-
Situation 67	FGMM	-	$\mathcal{M}_4$	Ξ	100	1.20	0	98.16	2.50	9	99.08	2.50	2	100	77.93	-

TABLE III
NORMAL OPERATING MODES OF CSTH

Case number	Mode label	Level SP	Temperature SP	Hot water valve
1	$\mathcal{M}_1$	11	9	4.5
	$\mathcal{M}_2$	12	9.5	4.5
	$\mathcal{M}_3$	12	11	5
	$\mathcal{M}_4$	9	10.5	4.5

TABLE IV
THE NUMBER OF REPLAY DATA

	Method	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
CSTH	MWCCA-A	20	30	36	30
	MNSDiPCA	95	86	73	71
TEP	MWCCA-A	20	29	16	24
	MNSDiPCA	1920	1920	1918	1919
Coal pulverizing system	MWCCA-A	34	62	19	62
	MNSDiPCA	36	319	67	481

data from different modes, which contributes to both reducing data imbalance and consolidating the significant features from previous modes. Consequently, although MWCCA-A uses fewer replay data than MNSDiPCA, MWCCA-A performs better on the previously learned modes than MNSDiPCA.

The offline computational complexity is measured by training time, as summarized in Table V. Note that mode information refers to the training data from these modes. As mentioned in Table II, training data are different for these comparative methods. For instance, different data are selected and replayed for MWCCA-A and MNSDiPCA. FGMM costs the most expensive computational resources. The complexity of MWCCA-A is less complicated than SDiPCA–MSI and FGMM. In contrast to MCVA and FGMM, the offline complexity of the four continual learning methods would not increase significantly with the successive emergence of new modes. According to aforementioned analysis, the proposed MWCCA-A approach delivers the best monitoring performance among the six methods, in terms of detection accuracy and computing costs.

B. Tennessee Eastman process

The Tennessee Eastman process is a complex industrial process and was widely utilized to illustrate the effectiveness of multimode monitoring methods [35], [36]. Detailed information was described in [37]. The data are collected from the Simulink model, which can be downloaded from http://depts.washington.edu/control/LARRY/TE/download.html. Four successive modes of process operation at three different G/H mass ratios are considered and listed in Table VI. In this experiment, 22 measured variables and 9 manipulated variables are utilized for monitoring. The sampling time is 3 minutes.

To construct the monitoring model, 1920 normal samples from each mode are collected. Two cases are considered, where testing samples are generated from two typical faults of TEP, namely, IDV(17) (Case 2) and IDV(19) (Case 3). Note that two cases share the same training data in Table VI. 1920 testing samples from each mode are collected, including the first 400 normal samples and subsequent 1520 faulty samples.

The monitoring results of Case 2 and Case 3 are listed in Table II. For two cases, the FDRs of MWCCA-A are higher than 97% and the FARs are lower than 6.8%. The FARs of Situations 10, 11 and 19 are higher than 15%, which reflects the catastrophic forgetting of WCCA-A for multiple mode problems. SDiPCA–MSI fails to monitor two cases accurately, because the FARs of Situations 24, 25 and 29 are higher than 18%. SPCA–SI could provide excellent performance, where the FDRs are higher than 94% and the FARs are not higher than 3.5%. MNSDiPCA offers higher accuracy than SPCA–SI, and the FDRs are higher than 97%. MCVA could not monitor Case 2 accurately, where the FDRs of Situations 51, 53, 54, 56–58 are lower than 90.5%. FGMM performs excellently on Cases 2 and 3 except Situations 62 and 65, where the FDRs are higher than 95% and the FARs are lower than 4.5%. For two cases, the FDRs of Situation 62 are lower than 6%, which indicate that FGMM fails to monitor the mode $\mathcal{M}_2$ based on the model $\mathcal{Z}$.

The selected data of MWCCA-A and MNSDiPCA are summarized in Table IV. Fewer than 2% of all normal training samples are selected for MWCCA-A. Significant weights are allocated to replay data in MWCCA-A, which makes it still perform excellently on the previous modes. MNSDiPCA and MWCCA-A provide similar monitoring performance with regard to Cases 2 and 3. However, almost all samples are selected based on cosine similarity and stored for MNSDiPCA. The training time is listed in Table V, where MWCCA-A is obviously less complicated than SDiPCA–MSI, SPCA–SI and MNSDiPCA. The complexity of MCVA is the lowest, but would increase with the sequential addition of dynamic modes. In conclusion, MWCCA-A provides the optimal performance among the six illustrative methods.

C. Coal pulverizing system

To demonstrate the effectiveness of MWCCA-A, a coal pulverizing system, which is one key unit of the 1030 MW ultra-supercritical thermal power plant in China, is employed. The structure and specific description can refer to [4], [12]. This work focuses on the abnormality from the coal feeder and detailed data information is summarized in Table VII. 14 critical variables have been pre-selected according to expert experience and prior knowledge.

The monitoring results are summarized in Table II. From Situations 1–19, it can be concluded that the proposed MWCCA-A method is able to monitor sequential modes accurately via a single model. The FARs of Situations 1 and 3 are 33.73% and 0.83%, which indicates that the information from the mode $\mathcal{M}_2$ is valuable to enhance the monitoring accuracy of the previous mode $\mathcal{M}_1$. Similar to CSTH case, the backward transfer learning ability of MWCCA-A is reflected, namely, the information from future modes contributes to boosting the performance of previous modes. The FARs of Situations 12 and 16 are 5.71% and 25.52%, which means that the significant information from the previous three modes, including the learned knowledge and partial representative data, is beneficial for enhancing the monitoring accuracy of the current mode $\mathcal{M}_4$. This phenomenon represents the forward

TABLE V
OFFLINE TRAINING TIME (s) OF ALL METHODS

Case number	Mode information	MWCCA-A	SDiPCA–MSI	SPCA–SI	MNSDiPCA	MCVA	FGMM
1	$\mathcal{M}_1$	45.11	67.63	13.49	6.05	-	-
	$\mathcal{M}_1, \mathcal{M}_2$	41.38	66.10	14.15	5.13	0.51	163.24
	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$	43.11	62.97	14.54	4.50	1.08	426.44
	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$	42.24	61.56	14.66	6.01	1.90	869.81
2 & 3	$\mathcal{M}_1$	209.71	2499.13	1369.56	4122.44	-	-
	$\mathcal{M}_1, \mathcal{M}_2$	193.91	2055.06	1316.93	4222.24	7.67	2111.01
	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$	208.67	1815.84	1435.55	4312.24	17.02	4879.32
	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$	209.64	1658.15	1437.45	4162.48	28.43	9494.71
4	$\mathcal{M}_1$	64.52	647.73	64.25	52.24	-	-
	$\mathcal{M}_1, \mathcal{M}_2$	121.38	1432.99	109.92	57.20	5.57	373.75
	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$	122.13	907.31	76.15	58.89	13.48	1077.22
	$\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$	119.79	1145.47	79.69	52.36	24.88	2267.44

TABLE VI
FOUR OPERATING MODES OF TEP

Case number	Mode label	Desired G/H mass ratio	Desired production
2 & 3	$\mathcal{M}_1$	50/50	14076
	$\mathcal{M}_2$	10/90	14077
	$\mathcal{M}_3$	90/10	11111
	$\mathcal{M}_4$	50/50	Maximum

transfer learning ability of continual learning. However, the FAR of Situation 11 is higher than 20%. The WCCA-A model again fails to monitor multiple modes and displays the catastrophic forgetting issue. Compared to the proposed method, the other five methods do not demonstrate excellent performance for sequential modes, where the FARs of several situations are especially high and are not acceptable. For instance, the FARs of Situations 22, 26, 34 and 36 are higher than 20%. In addition, the FARs of Situations 42, 44–47 are higher than 30%. For MCVA, the FARs of Situations 50, 52, 55, 57 and 58 are higher than 30%. For FGMM, the FARs are higher than 25% except for Situations 60, 62 and 65. In conclusion, the proposed method furnishes the highest monitoring accuracy among six methods.

The quantity of selected data needed to be stored for each model is summarized in Table IV. It is clear that MWCCA-A needs much less storage space than MNSDiPCA. With allocated weightings, the portion of data replayed in MWCCA-A can still reflect the operating condition of previous modes and then MWCCA-A delivers excellent performance on the previously learned modes. MWCCA-A only consumes a little more storage resources than SDiPCA–MSI and SPCA–SI. The training time is listed in Table V, the proposed MWCCA-A is less complicated than SDiPCA–MSI and FGMM. According to the analysis mentioned above, the proposed method provides optimal monitoring performance, in terms of accuracy and storage cost.

D. Ablation study

In this section, different data selection methods are considered and a relevant ablation study is conducted to illustrate the virtues of the proposed data selection method in Algorithm 1.

Various data selection methods have been investigated for replay continual learning. Gradient episodic memory selected representative data by minimizing negative backward transfer [38]. Gradient based sample selection (GSS) calculated the score of each sample based on the gradient and aimed to keep diverse samples in the replay buffer [39]. However, GSS requires that the gradient to be optimized should be calculated once for each iteration process, which makes it not appropriate within the framework of MWCCA-A. Adversarial Shapley value experience replay was investigated to score memory samples based on their ability to preserve latent decision boundaries [40]. It is extremely complicated because calculating the Shapley value requires $O(2^N)$ evaluations for general, bounded utility functions. The iCaRL method selected a subset of samples, to best approximate the average feature vector over all training examples [41]. It is computationally efficient and could order the importance of selected data.

In practical applications, it is important to select a simple and useful data selection method. Thus, random selection and iCaRL are adopted to conduct the ablation study. 40 representative samples are selected for replay continual learning. The hyperparameters are the same as those of the proposed MWCCA-A. The monitoring results are summarized in Table VIII, where only situations with regard to MWCCA-A are considered. To enhance the reliability of random selection, 200 independent repetition experiments are conducted and the average monitoring results are listed in Table VIII. For Case 1, the performance of random selection is not satisfactory because the FDRs are lower than 99%. Besides, the FARs of iCaRL are higher than those of the proposed method, where the FARs of Situations 13 and 14 are 5.8% and 4.8% respectively. With regard to Cases 2 and 3, the performance of random selection is excellent. However, the performance of iCaRL is not satisfactory because the FARs of Situation 8 are higher than 8% for both cases. For Case 4, the FAR of Situation 12 is 12.89% based on random selection. The FARs of Situations 8 and 12 are 10.75% and 19.80% using iCaRL.

According to the aforementioned analysis, the proposed data selection method based on k-means clustering and KNN is the optimal among three data selection methods, in consideration

TABLE VII
EXPERIMENTAL DATA OF THE PRACTICAL COAL PULVERIZING SYSTEM

Case number	Key variables	Mode label	Number of training data	Number of testing data	Fault location	Fault cause
4	14 variables: current and speed of coal feeder, rotary separator speed and current, coal feeding capacity, etc.	$\mathcal{M}_1$	1080	1080	846	The coal feeder does not drop coal
		$\mathcal{M}_2$	2160	1080	764	Speed probe failure
		$\mathcal{M}_3$	2160	1440	984	The coal feeder does not drop coal
		$\mathcal{M}_4$	2160	1440	1016	The coal feeder belt is broken

of complexity and monitoring performance.

VI. CONCLUSION

This paper has introduced the novel MWCCA-A method, which is a continual learning method using weighted CCA based on attention mechanism and aimed at multimode dynamic process monitoring. With dynamic, multimodal data being received sequentially, MWCCA-A works by extracting dynamic features via maximizing the weighted correlation between the latent variable and its prediction. Replay data from each mode are selected, and utilized in WCCA together with the current mode data in order to form a monitoring model with continual learning ability. To avoid potential data imbalance among different modes, the replayed data may be allocated large weighting and the significant features are consolidated further. Model parameter regularization is used based on synaptic intelligence. Finally, the effectiveness of the proposed MWCCA-A method is illustrated by CSTH, TEP and a practical coal pulverizing system. The experimental results have shown that the proposed approach outperforms several state-of-the-art multimode process monitoring methods, and is suitable for both long-term and short-term monitoring tasks.

In future, the CCA method within the framework of parameter isolation based continual learning would be investigated for multimode diverse dynamic modes.

REFERENCES

[1] Q. Wen, Z. Ge, and Z. Song, "Multimode dynamic process monitoring based on mixture canonical variate analysis model," *Industrial & Engineering Chemistry Research*, vol. 54, no. 5, pp. 1605–1614, 2015.

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

[3] R. Tan, J. R. Ottewill, and N. F. Thornhill, "Nonstationary discrete convolution kernel for multimodal process monitoring," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3670–3681, 2020.

[4] J. Zhang, D. Zhou, and M. Chen, "Monitoring multimode processes: a modified PCA algorithm with continual learning ability," *Journal of Process Control*, vol. 103, pp. 76–86, 2021.

[5] M. Quiones-Grueiro, A. Prieto-Moreno, C. Verde, and O. Llanes-Santiago, "Data-driven monitoring of multimode continuous processes: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 189, pp. 56–71, 2019.

[6] K. Huang, Z. Tao, Y. Liu, B. Sun, C. Yang, W. Gui, and S. Hu, "Adaptive multimode process monitoring based on mode-matching and similarity-preserving dictionary learning," *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3974–3987, 2023.

[7] Y. Dong and S. J. Qin, "A novel dynamic PCA algorithm for dynamic data modeling and process monitoring," *Journal of Process Control*, vol. 67, pp. 1–11, 2018.

[8] Y. Dong, Y. Liu, and S. J. Qin, "Efficient dynamic latent variable analysis for high-dimensional time series data," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4068–4076, 2019.

[9] Y. Dong and S. J. Qin, "Dynamic-inner canonical correlation and causality analysis for high dimensional time series data," *IFAC-Papers OnLine*, vol. 51, no. 18, pp. 476–481, 2018.

[10] X. Xu, J. Ding, Q. Liu, and T. Chai, "A novel multimanifold joint projections model for multimode process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 9, pp. 5961–5970, 2021.

[11] 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 Transactions on Automation Science and Engineering*, vol. 20, no. 1, pp. 137–150, 2023.

[12] J. Zhang, D. Zhou, and M. Chen, "Self-learning sparse PCA for multimode process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 29–39, 2023.

[13] J. Zhang, M. Chen, and X. Hong, "Monitoring multimode nonlinear dynamic processes: An efficient sparse dynamic approach with continual learning ability," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 7, pp. 8029–8038, 2023.

[14] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Network*, vol. 113, pp. 54–71, 2019.

[15] H. Li, P. Barnaghi, S. Enshaeifar, and F. Ganz, "Continual learning using Bayesian neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 4243–4252, 2020.

[16] G.-M. Park, S.-M. Yoo, and J.-H. Kim, "Convolutional neural network with developmental memory for continual learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2691–2705, 2020.

[17] A. Ororbia, A. Mali, C. L. Giles, and D. Kifer, "Continual learning of recurrent neural networks by locally aligning distributed representations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4267–4278, 2020.

[18] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022.

[19] A. Mallya and 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.

[20] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 3987–3995.

[21] E. S. Lubana, P. Trivedi, D. Koutra, and R. P. Dick, "How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation," *arXiv preprint arXiv:2102.02805*, 2021.

[22] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends Cognitive in Science*, vol. 24, no. 12, pp. 1028–1040, 2020.

[23] N. Y. Masse, G. D. Grant, and D. J. Freedman, "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. E10 467–E10 475, 2018.

[24] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *International Conference on Machine Learning*, 2018, pp. 4548–4557.

[25] J. Xu, J. Ma, and Z. Zhu, "Bayesian optimized continual learning with attention mechanism," *preprint arXiv:1905.03980*, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] X. Hong, J. Gao, and S. Chen, "Zero-attracting recursive least squares algorithms," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 213–221, 2017.

TABLE VIII
COMPARATIVE MONITORING RESULTS OF DIFFERENT DATA SELECTION METHODS

Methods		Model label	CSTH			TEP						Coal pulverizing system		
			Case 1			Case 2			Case 3			Case 4		
			FDR	FAR	DD	FDR	FAR	DD	FDR	FAR	DD	FDR	FAR	DD
Random selection	Situation 2	$\mathcal{B}$	85.87	0.45	0.31	96.57	0.83	21.11	99.87	0.83	1.275	100	0.27	0
	Situation 3	$\mathcal{B}$	89.81	0.53	1.22	98.96	1.92	15.625	99.48	1.92	4.25	100	0.50	0
	Situation 6	$\mathcal{E}$	92.89	0.44	26.50	98.66	0.13	15.35	99.85	0.13	2.10	100	0	0
	Situation 7	$\mathcal{E}$	97.08	0.61	0.98	98.96	2.08	14.60	99.60	2.09	4.31	100	0.52	0
	Situation 8	$\mathcal{E}$	95.54	0.49	0.145	96.30	4.99	21.105	99.88	4.99	0.835	100	0.32	0
	Situation 12	$\mathcal{F}$	96.37	0.51	1.065	98.99	0.20	4.75	99.49	0.21	0.105	100	12.89	0
	Situation 13	$\mathcal{G}$	98.78	1.72	0.07	98.97	2.31	15.44	99.60	2.31	4.25	100	0.39	0
	Situation 14	$\mathcal{G}$	98.26	0.55	0.01	96.41	2.13	20.935	99.88	2.14	1.085	100	0.26	0
	Situation 15	$\mathcal{G}$	96.96	0.46	0.165	98.84	3.52	13.24	99.89	3.52	1.54	100	0	0
iCaRL	Situation 2	$\mathcal{B}$	100	0.40	0	96.44	1.50	20	99.87	1.50	0	100	0.26	0
	Situation 3	$\mathcal{B}$	100	1.60	0	99.01	1.50	15	99.54	1.50	4	100	0.47	0
	Situation 6	$\mathcal{E}$	100	0.40	0	98.88	0	16	100	0	0	100	0	0
	Situation 7	$\mathcal{E}$	100	3.20	0	98.95	3.00	16	99.60	3.00	4	100	0.71	0
	Situation 8	$\mathcal{E}$	100	0.40	0	96.44	8.50	20	99.87	8.75	0	100	10.75	0
	Situation 12	$\mathcal{F}$	100	0.40	0	98.88	0.25	4	99.47	0.25	0	100	19.80	0
	Situation 13	$\mathcal{G}$	100	5.80	0	99.01	3.00	15	99.60	3.00	4	100	0.36	0
	Situation 14	$\mathcal{G}$	100	4.80	0	96.31	3.25	20	99.87	3.50	0	100	0.26	0
	Situation 15	$\mathcal{G}$	100	2.00	0	98.98	5.25	15	100	5.25	0	100	0	0

[28] S. Zhong, T. M. Khoshgoftaar, and N. Seliya, "Clustering-based network intrusion detection," *International Journal of Reliability Quality and Safety Engineering*, vol. 14, pp. 169–187, 2007.

[29] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[30] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley, 1992.

[31] T. Dozat, "Incorporating nesterov momentum into adam," *ICLR Workshop*, vol. 1, pp. 2013–2016, 2016.

[32] J. Yu and S. J. Qin, "Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models," *AIChE Journal*, vol. 54, no. 7, pp. 1811–1829, 2008.

[33] D. Cai, "Spectral regression: A regression framework for efficient regularized subspace learning," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2009.

[34] N. F. Thornhill, S. C. Patwardhan, and S. L. Shah, "A continuous stirred tank heater simulation model with applications," *Journal of Process Control*, vol. 18, no. 3, pp. 347–360, 2008.

[35] X. Xu, L. Xie, and S. Wang, "Multimode process monitoring with pca mixture model," *Computers & Electrical Engineering*, vol. 40, no. 7, pp. 2101–2112, 2014.

[36] B. Song, Y. Ma, and H. Shi, "Multimode process monitoring using improved dynamic neighborhood preserving embedding," *Chemometrics and Intelligent Laboratory Systems*, vol. 135, pp. 17–30, 2014.

[37] N.L.Ricker, "Optimal steady-state operation of the Tennessee Eastman challenge process," *Computer & Chemical Engineering*, vol. 19, no. 9, pp. 949–959, 1995.

[38] D. Lopez-Paz and M. A. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.

[39] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, "Gradient based sample selection for online continual learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 11 816–11 825.

[40] D. Shim, Z. Mai, J. Jeong, S. Sanner, H. Kim, and J. Jang, "Online class-incremental continual learning with adversarial shapley value," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9630–9638.

[41] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5533–5542.

Jingxin Zhang received B.E. degree in School of 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 lecture with the Department of Automation, Southeast University.

Her research interests are continual learning, data-driven fault detection and diagnosis, performance monitoring, photovoltaic power prediction and their applications in the industrial processes.

James Xiao received his MSci degree in Chemistry (2014) from the University of Bristol, UK, followed by his MRes in Nanoscience and Nanotechnology (2015) and PhD in Physics (2019) from the University of Cambridge, UK. From 2019 to 2022 he worked as a postdoctoral research associate at the Cavendish Laboratory, University of Cambridge. His current interests include machine learning and computer vision.

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.