Historical Information-aided Monitoring of Few-Sample Modes in Industrial Processes with Orthogonal Transferred Projection

Kai Wang, Xiang Lei, Wenxuan Zhou, Saige Cheng, and Jing Li

***Abstract*—Few-sample modes are easy to appear when a new working condition is triggered in industrial processes especially during the early stages of the new working mode. However, monitoring the early behaviour of a new mode is important because engineers and operators are less knowledgeable with such a new mode. Considering the few-sample challenge in this problem, a new multi-source transfer learning framework is proposed that leverages historical data under various operating conditions to enrich process monitoring over new mode data. In contrast to existing transfer learning-related work, a new unsupervised domain adaptation framework is designed. The historical modes as the source provide precious knowledge and reference to the new mode so that the features of the new mode are robust to noise and insufficient samples. Mathematically, the historical features play the role of a regularizer for the feature learning in the target domain. A geometrical illustration is given and an iterative optimization algorithm is developed with the convergence analysis. Except for the features guided by historical modes, individual features of the new mode are also extracted from the residual part to form a complete monitoring framework. Finally, the effectiveness of the proposed method is validated through a numerical experiment and a real industrial hydrocracking process.**

***Index Terms*—Process monitoring, multimode data, few samples, transfer learning, common subspace**

I. Introduction

In modern industrial processes, process industry is constantly developing towards integration, large-scale, and intelligence [1], [2]. Its characteristics, including long process flows, numerous equipment, and complex mechanism, often make it difficult to apply traditional process monitoring methods, such as mechanism-based or knowledge-based approaches. Recently, the increasing use of distributed control systems and industrial internet enables access to large volumes of data, making data-driven process monitoring technology a current research topic [3]. Typically, process information can be mined through historical process data upon which a mathematical description of the process can be established. For instance, multivariate statistical analysis methods such as principal component analysis (PCA) [4] and partial least squares [5] have been successfully used in operational assessment and fault detection/diagnosis.

Recent developments of multivariate statistical process monitoring are towards the direction of how to improve the monitoring performance or provide additional information. The structured PCA with variable correlation constraint is proposed for embedding a priori knowledge and making the data-driven model more interpretable [6]. Recursive correlative statistical analysis by analyzing the correlation within a sliding window is proposed for detecting the incipient fault [7]. Robust and sparse CCA could train the data with outliers and separate the noise from data well [8]. To deal with multirate sample problem, Zhou et al. designed a multirate mixture probability PCA for sufficiently use the collected data [9]. However, data-driven methods depend on sufficient data. Otherwise, the data model would lose its representative ability for the real process. In a real process, the fact is that operating conditions could frequently change due to various practical factors [2], such as raw material fluctuations, personnel operations, instrument failures, market changes, etc. The multimodal behavior has become a fundamental characteristic in the historical process profile. However, not all newly triggered working conditions are repetition of historical modes. Especially, new working conditions frequently appear in some newly built plants. Moreover, practitioners generally learn less about the new mode than those historical repetitive modes. Hence, anomaly can occur in the initial stage of the new mode with a higher risk. This stimulates a strong demand for monitoring the initial stage with few samples. However, the few samples has been the most challenging obstacle for monitoring the early stage of a new mode. This is because the monitoring performance is dependent on the extraction of process correlations. However, due to noise commonly existing in industrial systems, the quality of process models built by insufficient samples is really poor. The noise will heavily influence the identification of process correlations.

Few-sample learning, also known as few-shot tasks, has received much attention in recent researches. Data augmentation could be a direct idea for augmenting the training set, which generally adapt the existing dataset to more generalized datasets. This is achieved by simple data transformation [10] or learning a generative model [11], [12]. However, data

This work was supported in part by the National Natural Science Foundation of China (62373378), in part by the Natural Science Foundation of Hunan Province in China (2022JJ20079), and in part by the Central South University Innovation-Driven Research Program, Changsha in China (2023CXQD054). The work of Jing Li was supported by Maritime AI Research Programme of Singapore under Grant SMI-2022-MTP-06.

K. Wang, X. Lei, W. Zhou, and S. Chen are with the School of Automation, Central South University, Changsha 410083, China (kaiwang@csu.edu.cn, xianglei@csu.edu.cn, zhouwenxuan1@csu.edu.cn, sgcheng@csu.edu.cn).

Jing Li is with the Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore and the Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore 138632. (li_jing@cfar.a-star.edu.sg).

Corresponding authors: Wenxuan Zhou and Jing Li.

augmentation often under-performs in face of few samples because few samples are non-representative for the genuine features [13]. On the one hand, the data in industrial processes is unlike images which could be augmented by scaling, rotation, translation or noise enhancement. This actually supplies additional information for learning the model. In other words, in the population, there exists these transformed images while they don't be sampled. Industrial data consist of temperature, flow rate, pressure and other process variables with process noise. In face of few sample issues, we could indeed consider sample generation strategies for increasing the amount of data. However, it is hard to evaluate whether the generated samples are beneficial for estimating the true data distribution. Especially, the virtual datapoints are generated by the collected samples, in which no new information is introduced. Since the crucial idea is how to leverage existing information into the few-sample task. Meta-learning resorts to the models trained by similar tasks and transfer the model into the new few-sample task [14]. Hence, it is a task-level optimization method such as the model-agnostic meta-Learning that seeks an optimal initial states from other similar tasks to drive the new meta-learner [15]. It needs a large number of learning tasks to help determine the suitable hyper-parameters and initialized parameters for the new task. Technically, the few samples make a function of fine-tuning the parameters for better adapting to the new task. For the industrial processes, preparing a large number of tasks based on historical data is very cost. Moreover, there is little research on the meta-learning-based industrial process monitoring. Domain adaption that could be understood as the feature-level or data-level knowledge transfer is more suitable to industrial process monitoring [16]. Zhang et al. [17] proposed a domain discrepancy-guided contrastive feature learning framework for few-shot fault diagnosis under varied working conditions. Moreover, a dual graph neural network is designed for improving the diagnosis precision for mechanical facilities under few-sample conditions [18].

Even though there exists advanced few-shot learning methods, some special requirements should be considered in industrial process monitoring. Firstly, with available process variables, monitoring models are learned in a manner of unsupervision. This is actually different from most of the paradigm of few-sample learning or transfer learning, which generally implements feature alignment between the source domain and the target domain for improving prediction performance. Under the label guidance, the features in the source domain and the target domain are aligned towards the optimal prediction performance of the target domain. However, just mixing the source features and the target features is not necessarily beneficial for improving monitoring performance. There is no theoritical guarantee that the data in the future can be mapped into the trained feature space because the few samples cannot represent the whole distribution in the target domain. Even though there exists related work for unsuperivsed transfer learning, they are mostly based on deep learning that lack good interpretability for which industrial processes poses a strong requirement because the running safety is the fundamental condition in industrial systems [19]. Moreover, even though these methods claim an unsupervised transfer paradigm, they actually still assume one of domains is supervised, or a supervised task is achieved such as classification or prediction [20]. Process monitoring tasks are actually different from the prediction or classification. It is a totally behaviour of the unsupervised learning and aims to learn data distribution and the boundaries. Therefore, few-sample process monitoring concerns more about how the historical knowledge can be utilized and what kind of historical knowledge should be transferred in the situation of insufficient data. Then, orthogonal projection property is also important for splitting the original data space into several subspace because monitoring results in each subspace can provide important anomaly information.

Notice the new mode can appear in the running process while not all process correlations are changed and the degree of change is limited because the process running observes the constraints of real physical and chemical laws [21]. For example, nonlinearities are common in complex industrial equipment. Because the process fluctuation is generally around a working point, linear model makes sense given a single mode. It is easy to know the linearized parameters can be different between modes while the model structures and parts of parameters may be very similar. This means the new mode can be different from but similar as historical data, which is the base of reusing historical knowledge. Now the problem is that what kind of knowledge should be leveraged into the new mode. Without any prior information, the most robust strategy is that selecting the features shared by nearly all historical modes to transfer. Then, it is of high likelihood that these common features are contained in the new mode, which is like an average operation. Moreover, monitoring the common subspace and the individual subspace between multiple operating conditions is useful for locating the root cause of failure and has become a new paradigm of process monitoring.

In the literature, various methods have been proposed to obtain common features. One group considers that the common score matrix can be derived by weighting the score matrix of each mode [22]. In [23], the two-step multi-set basis vector extraction algorithm is proposed where basis vectors are constructed to describe cross-set correlation. And the other group method is to derive the auxiliary common scores from individual features of each mode, and then direct mapping of each mode to auxiliary common features which are used for common feature extraction. Thus they are also viewed as two-stage methods for extracting common components [24]. Additionally, there are other works that consider both common scores and common loadings simultaneously, e.g., tensor decomposition-based methods. Zhang et al. [21] proposed a canonical polyadic decomposition (CPD)-based method for extracting common features. Liu et al. [25] proposed the Tucker decomposition-based supervision method to calculate the common features and applied in the multi-grade process. In the field of tensor decomposition, CPD is more widely used due to its high interpretability. Thanks to this advantage, CPD-based methods have also been widely used in field of image and pattern recognition [26].

However, these common components are extracted and

utilized for monitoring historical repetitive models rather than the unseen new mode. In our paper, the target domain refers to the new operating conditions with few samples [27]. In the early stage of switching of working conditions, there will be a cold start problem that the amount of data is insufficient, which makes it difficult to build a model. Transfer learning can effectively guide the modeling of new operating conditions by excavating the domain-invariant features in historical multimode data. The problem of model overfitting caused by insufficient data is thereby alleviated [2]. Since the problem in this paper is a totally unsupervised learning, it is important figuring out what kind of historical knowledge should be transferred and how to transfer it. To this aim, we stack historical multimode data into a three-way tensor form beforehand. And we select CPD-based method to extract common features from multimodal data. Further a reasonable interpretation is given for the function of the common features. Procrustes analysis is used to align the common features of the source and target domains, then, the objective function is solved by using the Orthogonal Procrustes analysis to achieve knowledge transfer. Moreover, to further get the individual features of the target domain data, we decompose the residual part of the target domain to obtain the individual feature part and noise part of the target domain. The problem to be solved and the main contributions of this paper are summarized as follows:

- The monitoring problem in the new mode at the early stage is considered rather than the historically repetitive working modes. In the early stage of a new mode, the samples collected are insufficient, which poses a dominant challenge on data modelling and monitoring. The transfer learning that leverages historical similar information is considered to solve such a problem.
- For effectively transferring the historical information into the new mode for a monitoring task, the orthogonal components when extracting features of the new mode are guaranteed, avoiding the extraction of redundant information. Moreover, a geometrical interpretation is given for why the orthogonal component extraction strategy is effective. Based on the geometrical interpretation, the suitable application scenarios are also illustrated.
- The proposed feature extraction method is achieved by an iterative optimization algorithm. The convergence is guaranteed through algorithm analysis. Moreover, the model proposed is parametric so that the computational complexity is very low for online monitoring.

This paper is organized as follows. Section II introduces the method of CPD to extract the common component and the strategy to transfer the common component. The solution procedure of the transfer learning model is given in Section III. The Section IV calculates the statistics and corresponding control limits of common part, individual part and noise part. The Section V verifies the effectiveness of the method with a numerical example and a hydrocracking example. Section VI concludes the paper.

II. MULTI-SOURCE TRANSFER LEARNING FORMULATION

In this section, we formulize the transferring learning problem using the multi-source historical data. To effectively compress the large-scale historical data and extract historical useful information, the tensor decomposition-based CPD is firstly introduced for extracting historical mode-invariant components which. Then, the model to be learned for monitoring the new mode, which should incorporate the historical mode-invariant components, is constructed. Moreover, to avoid the information redundancy for the extracted features, the orthogonality constraint of features is designed and embedded into the proposed model, which help improve the monitoring performance. Since the learning for the model is totally unsupervised, it is not easy to understand why the transferring is effective because there is no label guidance. Therefore, a novel geometrical interpretation is presented for illustrating the learning process.

A. Common component extraction

In most of industrial processes, there exists part crucial variables that always are maintained within the same safe scope cross different kinds of working conditions. Part correlations between variables are also similar under different modes. This means the coming new mode will observe such constant rules. To monitor such a new mode where just few samples are collected in the beginning, the invariant features become so valuable knowledge for judging process behavior. As a result, the CPD-based shared feature extraction cross multimode data is employed first to extract the common components. And domain adaptation will leverage the common feature to the new mode.

Specifically, given a multimode data $\underline{\mathbf{X}} \in \mathbb{R}^{I\times J\times K}$ that are three-dimensional and illustrated in Fig. 1. Here I is the number of samples, J is the number of variables for each mode and K is the number of modes. $\underline{\mathbf{X}}$ can be decomposed as $\underline{\mathbf{X}} = \underline{\mathbf{X}}_c + \underline{\mathbf{E}}_c$, where $\underline{\mathbf{X}}_c$ and $\underline{\mathbf{E}}_c$ are common part and common residual part respectively.

Through the CPD-based method, the common score matrix and common load matrix in the process data of multiple working conditions can be obtained simultaneously. Since $\underline{\mathbf{X}}$ is a tensor, it can be decomposed as

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \underline{\mathbf{X}}_r = \sum_{r=1}^{R} \boldsymbol{t}_r \circ \boldsymbol{p}_r \circ \boldsymbol{w}_r \tag{1}$$

where the symbol $\circ$ is the vector outer product, R is the rank of tensor, $R \leq \min\{IJ, IK, JK\}$. This approximation expresses $\underline{\mathbf{X}}$ as the sum of rank-1 tensors $\underline{\mathbf{X}}_r$, and $\boldsymbol{t}_r \in \mathbb{R}^I$, $\boldsymbol{p}_r \in \mathbb{R}^J$ and $\boldsymbol{w}_r \in \mathbb{R}^K$ are called as factorized vectors [21], [28]. Likewise, there is

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \underline{\mathbf{X}}_{c,r} = \sum_{r=1}^{R} \boldsymbol{t}_{c,r} \circ \boldsymbol{p}_{c,r} \circ \boldsymbol{w}_{c,r} \tag{2}$$

One can see all these factors form corresponding matrices given by $\mathbf{T}_c = [\boldsymbol{t}_{c,1}, \boldsymbol{t}_{c,2}, \cdots, \boldsymbol{t}_{c,R}] \in \mathbb{R}^{I\times R}$, $\mathbf{P}_c = [\boldsymbol{p}_{c,1}, \boldsymbol{p}_{c,2}, \cdots, \boldsymbol{p}_{c,R}] \in \mathbb{R}^{J\times R}$ and $\mathbf{W}_c = [\boldsymbol{w}_{c,1}, \boldsymbol{w}_{c,2}, \cdots, \boldsymbol{w}_{c,R}] \in \mathbb{R}^{K\times R}$, which can be regarded as

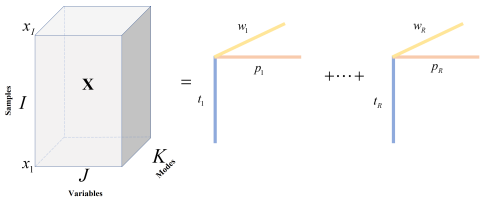


Fig. 1: CP decomposition method.

matrix forms of common features embedded in tensor $\underline{\mathbf{X}}$. To learn these common features, we should unfold the tensor into matrices. Specifically, $\underline{\mathbf{X}}_c$ can be unfolded along the mode, i.e., $[\mathbf{X}_{[1],c}, \mathbf{X}_{[2],c}, \cdots, \mathbf{X}_{[k],c}, \cdots, \mathbf{X}_{[K],c}]$. There is

$$\mathbf{X}_{[k],c} = \sum_{r=1}^{R} \boldsymbol{t}_{c,r} \boldsymbol{p}_{c,r}^T \boldsymbol{w}_{[k],c,r} = \mathbf{T}_c \mathbf{W}_c^{(k)} \mathbf{P}_c^T \tag{3}$$

where $\mathbf{W}_c^{(k)} = diag\left(\boldsymbol{w}_{[k],c}\right) \in {}^{R\times R}$, with $k = 1, 2, \cdots, K$. $\mathbf{T}_c$ and $\mathbf{P}_c$ are the common scores and loadings, analog to PCA.

Moreover, the common features should minimize the reconstruction errors for the original observations. Notice all the three matrices in (3) are unknown, alternating least squares (ALS) can be used to implement reconstruction error minimization, given by

$$\begin{cases} \min\limits_{\mathbf{T}_c} \left\| \mathbf{X}_{(1)} - \mathbf{T}_c (\mathbf{W}_c \odot \mathbf{P}_c)^T \right\|_F^2 \\ \min\limits_{\mathbf{P}_c} \left\| \mathbf{X}_{(2)} - \mathbf{P}_c (\mathbf{W}_c \odot \mathbf{T}_c)^T \right\|_F^2 \\ \min\limits_{\mathbf{W}_c} \left\| \mathbf{X}_{(3)} - \mathbf{W}_c (\mathbf{P}_c \odot \mathbf{T}_c)^T \right\|_F^2 \end{cases} \tag{4}$$

where $\|\cdot\|_F^2$ is the Frobenius norm, $\odot$ is the Khatri-Rao product, $\mathbf{X}_{(i)}$ is the unfolding of $\underline{\mathbf{X}}$ along three different dimensions.

When the iteration converges, $\mathbf{T}_c$, $\mathbf{P}_c$ and $\mathbf{W_c}$ can be derived as

$$\mathbf{T}_c = \mathbf{X}_{(1)} \left[(\mathbf{W}_c \odot \mathbf{P}_c)^T \right]^{\dagger} \tag{5}$$

$$\mathbf{P}_c = \mathbf{X}_{(2)} \left[(\mathbf{W}_c \odot \mathbf{T}_c)^T \right]^{\dagger} \tag{6}$$

$$\mathbf{W_c} = \mathbf{X}_{(3)} \left[(\mathbf{P}_c \odot \mathbf{T}_c)^T \right]^{\dagger} \tag{7}$$

Following the extraction of the common features of historical multimode data as (5)-(7), the common features of the new working condition need to be transferred from the common features of the historical multimode data.

B. Feature direction transfer

When the new mode data $\mathbf{X}^{TD}$ generate, to implement effective knowledge transfer, it is necessary to figure out what should be transferred. Since the focus is boosting the monitoring performance in the new mode, so-called feature alignment is not adopted because we cannot guarantee the samples in the coming future also follow the same distribution with the mixed features. Instead , it is approved just part variable correlations are changed under different running modes, which means shared variable correlations are crucial knowledge to be leveraged. Notice $\mathbf{X}_{(2)}$ is the variable-wise unfolded matrix. Thus 6 is the shared variable correlations observed by historical datasets. This can also be observed by 3 which symbolizes the generative model that generates $\mathbf{X}_{[\mathbf{k}],\mathbf{c}}$ from $\mathbf{T}_c \mathbf{W}_c^{(k)}$. Therefore, the common loadings $\mathbf{P}_c$ that indicates how the variable correlates with each other are suitable knowledge to be transferred. The remaining problem is how to reuse $\mathbf{P}_c$ in the new mode.

Multivariate process monitoring prefers the property of orthogonal subspace decomposition for detecting the anomaly in different subspaces, which can revealing parts of fault attributes. However, it is easy to know the procedure of CPD that learns $\mathbf{P}_c$ does not guarantee the orthogonality. Nonetheless, the orthogonality of the target domain can be sought due to the focus on monitoring performance in the new mode. Denote the learned common loadings as $\mathbf{P}_{c,s}$ and the loading to be learned in the target domain as $\mathbf{P}_{c,t}$. For reasonable adaptation of an orthonormal matrix to a non-orthonormal matrix. We need a translation-dilation-rotation transformation that does not change the intrinsic correlations of variables to transform an orthonormal matrix to an ordinary matrix as shown in Fig. 3. Therefore, the alignment of loadings between the source domain and the target domain is defined as

$$\begin{aligned} &\arg\min_{\rho,\mathbf{Q}} \|\mathbf{P}_{c,s} - (\rho \mathbf{P}_{c,t} \mathbf{Q} + \boldsymbol{b}_k)\|_F^2 \\ &s.t. \mathbf{Q}^T \mathbf{Q} = \mathbf{I} \end{aligned} \tag{8}$$

Because coordinate rotation does not change the process correlations, an orthonormal matrix $\mathbf{Q}$ is added to allow the rotation. ρ is a dilation factor to adjust the vector norm. $\boldsymbol{b}_k$ is the centralized translation matrix. Note that the translation term can be set to zero if the subspaces of both the source and target domains have been moved to the origin [29]. Therefore, $\mathbf{P}_{c,t}$, $\mathbf{Q}$ and ρ are three parameters to be solved.

Besides the alignment objective, the process structure partly unveiled by the few samples in the new mode is also expected to be learned. Otherwise, the negative transfer can be easily triggered because the adaptation does not represent any information of the new mode. Thus, the whole objective of the multi-source common feature transfer model is

$$\begin{aligned} &\arg\min_{\mathbf{T}_{c,t},\mathbf{P}_{c,t}} \left\| \mathbf{X}^{TD} - \mathbf{T}_{c,t} (\mathbf{P}_{c,t} \mathbf{Q})^T \right\|_F^2 + \lambda \left\| \mathbf{P}_{c,s} - \rho \mathbf{P}_{c,t} \mathbf{Q} \right\|_F^2 \\ &s.t. \mathbf{P}_{c,t}^T \mathbf{P}_{c,t} = \mathbf{I}, \mathbf{Q}^T \mathbf{Q} = \mathbf{I} \end{aligned} \tag{9}$$

where $\mathbf{T}_{c,t} \in \mathbb{R}^{I_t \times R}$ is the score matrix of the target domain model and is the balance coefficient. One can see the source data do not participate in the transfer while the feature direction learned by the source data is aligned. This is different from most of domain adaption methods. Moreover, (9) can be regarded as an Orthogonal Procrustes problem [30], which will be briefly introduced in next section. Note that the $\mathbf{X}^{TD}$ needs normalization before optimizing (9) .

C. Alignment-based Regularization

In this subsection, we will give an illustration of the working principle of the loading alignment in (9). Intrinsically, (9)

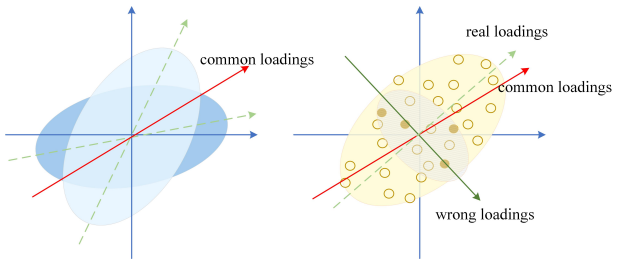


Fig. 2: Geometrical illustration.

indicates the feature learning in the new mode is regularized by the alignment of the source and the target feature directions. Due to few samples in the target mode, an intuition is that the loading adaption in 8 could make sense for anti-overfitting or anti-singularity. However, the model structure is simply linear and there exists the orthogonal constraints for $\mathbf{P}_{c,t}$, which means no strong demands may be imposed on the feature learning. Fig. 2 reveals why such a regularization is indispensable in few samples and what kind of historical knowledge is inherited by the new mode. Specifically, the two variables are correlated and there exists a direction representing the correlation. As shown in the left of Fig. 2, the loading directions are different but close to each other in the two historical modes. Using tensor decomposition in (6), the 'averaged' common loadings marked in red can be obtained. One can see the projection on the common loadings can still capture the dominant process fluctuation, though the fluctuation extracted is not the largest variance. When it comes to the new mode in the right of Fig. 2, there are just four samples marked in solid circles collected as the new mode just start to run. As the running process progresses, more samples could be collected in the future, marked in hollow circles. When enough samples are available, process features can be learned well without using any historical information. However, at the early stage, these hollow circles are unknown and one can see the loadings learned by just few samples can even be orthogonal to the true loadings, which could completely mislead engineers to conduct a wrong conclusion. Instead, if the historical loadings can be aided to learn the features in the new mode, the false direction can be alleviated a lot.

Fig. 2 also implies the probable application situations in industrial systems, which requires the historical modes are similar and the new mode is similar as the historical modes, as well. This can avoid the negative transfer to a great extent. As aforementioned, most of running modes are different in working points so that the variable correlations are similar. This characteristic renders the proposed method a suitable application situation. As the data scale in the new mode increases, just using the newly collected data could be better because the historical common loadings are not completely consistent with the genuine loadings of the new mode. Compulsory transfer will cause severe side effect and diminish the ability of feature learning.

III. SOLUTION FOR THE TRANSFER LEARNING MODEL

It is clear that (9) is a bilinear optimization problem with orthonormal constraints. The optimization step of the algorithm can be divided into two steps. Solve (8) and (9) through Step 1 and Step 2 respectively. Specifically, use PCA to initialize $\mathbf{P}_{c,t}$ and $\mathbf{T}_{c,t}$ by decomposing $\mathbf{X}^{TD}$.

Step 1: Update the rotation matrix $\mathbf{Q}$, dilation coefficient ρ given $\mathbf{P}_{c,t}$ and $\mathbf{T}_{c,t}$. In this step, the objective is completely reduced to (8). As aforementioned, (8) is a matrix trace minimization problem with orthonormal constraints, which can resort to Procrustes similarity analysis [31] for optimal solution. Specifically, performing singular value decomposition yields $\mathbf{P}_{c,s}^T\mathbf{P}_{c,t} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, then there is $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$ and further $\rho = trace\left(\mathbf{P}_{c,s}\mathbf{Q}^T\mathbf{P}_{c,t}^T\right)/trace\left(\mathbf{P}_{c,t}\mathbf{P}_{c,t}^T\right)$.

Step 2: With the iterative dilation factor ρ and the coordinate rotation matrix $\mathbf{Q}$, it is necessary to update the common loading matrix $\mathbf{P}_{c,t}$ and core matrix $\mathbf{T}_{c,t}$ in the target domain. By introducing data $\mathbf{X}^{TD}$ from the target domain, the reconstruction loss of the common subspace is calculated. Then (9) is converted to another orthogonal Procrustes problem, which has the following solutions.

$$\begin{aligned}&\arg\min_{\mathbf{T}_{c,t},\mathbf{P}_{c,t}} \left\|\mathbf{X}^{TD} - \mathbf{T}_{c,t}(\mathbf{P}_{c,t}\mathbf{Q})^T\right\|_F^2 + \lambda\left\|\mathbf{P}_{c,s} - \rho\mathbf{P}_{c,t}\mathbf{Q}\right\|_F^2\\ \Rightarrow &\arg\max_{\mathbf{T}_{c,t},\mathbf{P}_{c,t}} trace\left(\mathbf{P}_{c,t}\mathbf{Q}\mathbf{T}_{c,t}^T\mathbf{X}^{TD}\right) + \lambda trace\left(\rho\mathbf{P}_{c,t}\mathbf{Q}\mathbf{P}_{c,s}^T\right)\\ &s.t.\mathbf{P}_{c,t}^T\mathbf{P}_{c,t} = \mathbf{I}, \mathbf{Q}^T\mathbf{Q} = \mathbf{I}\end{aligned} \tag{10}$$

Since the optimization problem is not a joint convex problem for both $\mathbf{P}_{c,t}$ and $\mathbf{T}_{c,t}$, but is separately convex to $\mathbf{P}_{c,t}$ (while holding $\mathbf{T}_{c,t}$ fixed) and convex to $\mathbf{T}_{c,t}$ (while holding $\mathbf{P}_{c,t}$ fixed) [32]. Therefore, an iterative optimization method is introduced to obtain the optimal values of $\mathbf{P}_{c,t}$ and $\mathbf{T}_{c,t}$.

First, update $\mathbf{P}_{c,t}$ while holding $\mathbf{T}_{c,t}$ fixed. Notice (10) aims to maximize the trace of the matrix $\mathbf{P}_{c,t}\mathbf{W}$ where $\mathbf{W}$ is $\mathbf{Q}\mathbf{T}_{c,t}^T\mathbf{X}^{TD} + \lambda\rho\mathbf{Q}\mathbf{P}_{c,s}^T$. Perform SVD on $\mathbf{W}$, i.e., $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. There is further

$$\begin{aligned}&\arg\max_{\mathbf{P}_{c,t}} trace\left(\mathbf{P}_{c,t}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right) = trace\left(\mathbf{\Lambda}\mathbf{V}^T\mathbf{P}_{c,t}\mathbf{U}\right)\\ &s.t.\mathbf{P}_{c,t}^T\mathbf{P}_{c,t} = \mathbf{I}, \mathbf{Q}^T\mathbf{Q} = \mathbf{I}\end{aligned} \tag{11}$$

Since $\mathbf{\Lambda}$ is diagonal, there is

$$trace\left(\mathbf{\Lambda}\mathbf{V}^T\mathbf{P}_{c,t}\mathbf{U}\right) = \sum_i \Lambda_{i,i}\left(\mathbf{V}^T\mathbf{P}_{c,t}\mathbf{U}\right)_{i,i} \tag{12}$$

where the subscript i,i denotes the i-th diagonal element of a matrix. We can see $\mathbf{V}$, $\mathbf{P_{c,t}}$ and $\mathbf{U}$ are all orthonormal matrices. The diagonal element of their product $\left(\mathbf{V}^T\mathbf{P}_{c,t}\mathbf{U}\right)_{i,i}$ should be no more than 1, i.e.,

$$\sum_i \Lambda_{i,i}\left(\mathbf{V}^T\mathbf{P}_{c,t}\mathbf{U}\right)_{i,i} \le \sum_i \Lambda_{i,i}\mathbf{I}_{i,i} = \sum_i \Lambda_{i,i} \tag{13}$$

(13) take the equal sign if and only if $\mathbf{V}^T\mathbf{P}_{c,t}\mathbf{U} = \mathbf{I}$ so that $\mathbf{P}_{c,t} = \mathbf{V}\mathbf{U}^T$ is the solution of (11). Finally, the loadings learned are $\mathbf{P}_{c,t} = \mathbf{P}_{c,t}\mathbf{Q} = \mathbf{V}\mathbf{U}^T\mathbf{Q}$. Sequentially, the score matrix is updated as $\mathbf{T}_{c,t} = \mathbf{X}^{TD}\mathbf{P}_{c,t}$ while holding $\mathbf{P}_{c,t}$ fixed.

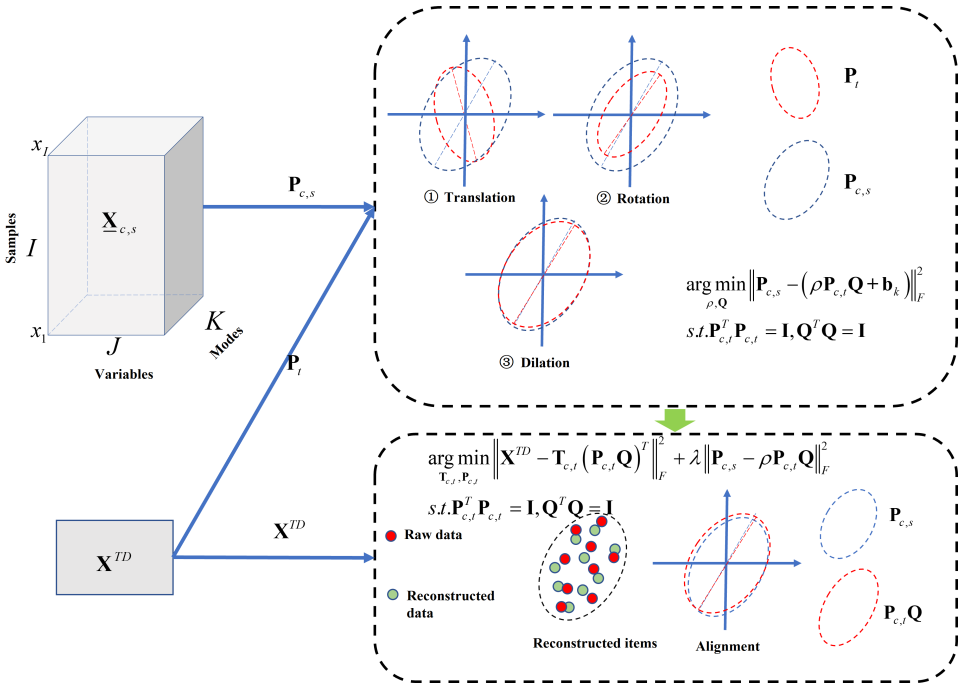


Fig. 3: Multi-source common feature transfer learning.

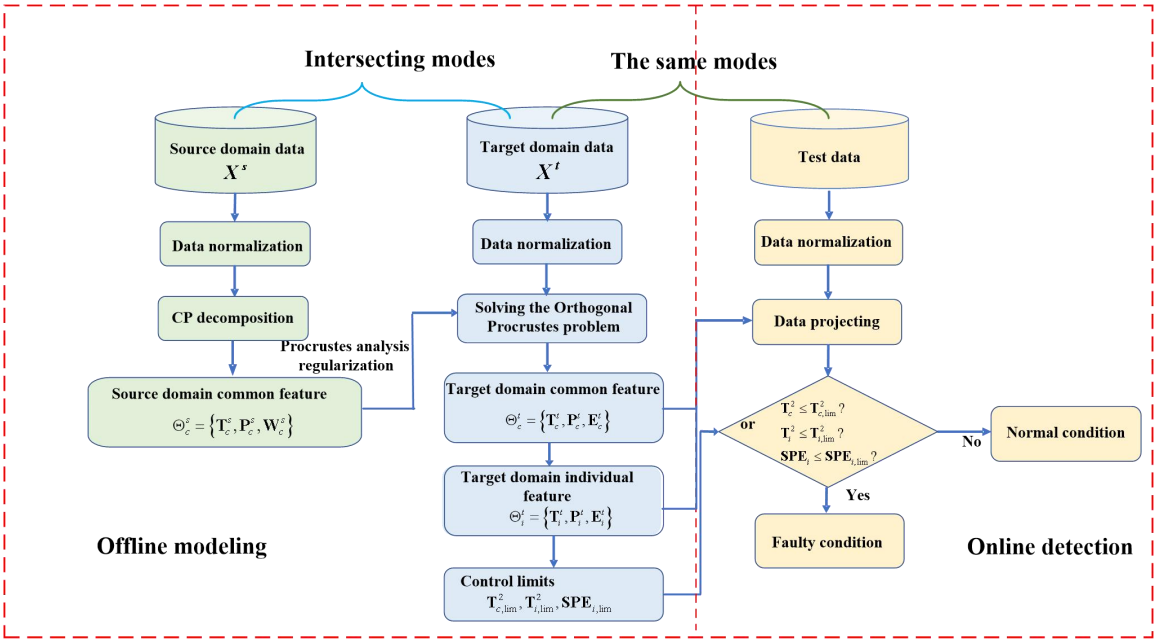


Fig. 4: MsTL fault detection flowchart for offline modeling and online fault detection.

To sum up, the proposed optimization issue is divided into two steps and further transform into separately convex problems. The algorithm will achieve a smaller objective value at each iteration so that the algorithm can be finally converged. Then the iteration termination condition can be defined by the maximum number of iterations or the objective evolving curve. The optimization procedure is shown in Algorithm 1. After completing the optimization steps, the common part of the target domain can be obtained, i.e., $\mathbf{X}_{c,t} = \mathbf{T}_{c,t}\mathbf{P}_{c,t}^{T}$.

IV. APPLICATION TO PROCESS MONITORING

The fault detection indices should be built for online monitoring after the multi-source transfer learning model is trained. Notice we here pursue satisfactory knowledge transfer rather than information preservation. Therefore, besides the common components extracted by solving the transfer learning problem, part information representing the fluctuation of the target mode is contained in the individual components $\mathbf{E}_{c,t}$ and there is

$$\mathbf{E}_{c,t} = \mathbf{X}^{TD} - \mathbf{X}_{c,t} \tag{14}$$

Algorithm 1 MsTL Optimization Procedure.

Input: Target domain data $\mathbf{X}^{TD}$, common loadings $\mathbf{P}_{c,s}$ in source domain, maximum number of iterations l

Output: Common loadings matrix $\mathbf{P}_{c,t}$, score matrix $\mathbf{T}_{c,t}$ in target domain

Initialization: use PCA to initialize $\mathbf{P}_{c,t}$ and $\mathbf{T}_{c,t}$ by decomposing $\mathbf{X}^{TD}$;
for $i = 1, 2..., l$ **do**
  // Step 1
  $\mathbf{P}_{c,s}^T \mathbf{P}_{c,t} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$;
  Update the rotation matrix: $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$;
  Update the dilation coefficient: $\rho = trace\left(\mathbf{P}_{c,s}\mathbf{Q}^T\mathbf{P}_{c,t}^T\right)/trace\left(\mathbf{P}_{c,t}\mathbf{P}_{c,t}^T\right)$;
  // Step 2
  $\mathbf{Q}\mathbf{T}_{c,t}^T\mathbf{X}^{TD} + \lambda\rho\mathbf{Q}\mathbf{P}_{c,s}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$
  Update the common loadings: $\mathbf{P}_{c,t} = \mathbf{V}\mathbf{U}^T\mathbf{Q}$
  Update the score matrix: $\mathbf{T}_{c,t} = \mathbf{X}^{TD}\mathbf{P}_{c,t}$
end for

In this case, the detection indices can be designed in the common component subspace and the individual component subspace, respectively.

Through PCA, it is easy to decompose $\mathbf{E}_{c,t}$ into the principal component subspace and the residual subspace, respectively, given by

$$\mathbf{E}_{c,t} = \mathbf{X}_{i,t} + \mathbf{E}_{i,t} = \mathbf{T}_{i,t}\mathbf{P}_{i,t}^T + \mathbf{E}_{i,t} \tag{15}$$

In this case, cross validation and cumulative percentage variance (CPV) can be utilized to calculate the number of principal components of each mode [22].

A sample from new target domain arrives, the common score is

$$\boldsymbol{t}_{new,c,t} = \mathbf{P}_{c,t}^T\boldsymbol{x}_{new,t} \tag{16}$$

Correspondingly, the residual part of $\boldsymbol{x}_{new,t}$ is $\boldsymbol{e}_{new,c,t} = \left(\mathbf{I} - \mathbf{P}_{c,t,new}\mathbf{P}_{c,t,new}^T\right)\boldsymbol{x}_{new,t}$. Similarly, for individual components, there are

$$\boldsymbol{t}_{new,i,t} = \mathbf{P}_{i,t}^T\boldsymbol{e}_{new,c,t} \tag{17}$$

$$\boldsymbol{e}_{new,i,t} = \left(\mathbf{I} - \mathbf{P}_{i,t}\mathbf{P}_{i,t}^T\right)\boldsymbol{e}_{new,c,t} \tag{18}$$

The Hotelling's T^2 statistic to measure the variations in the common subspace and the principal component subspace of individual subspace is constructed as follows

$$T_{new,c}^2 = \boldsymbol{t}_{new,c,t}^T\left(\frac{1}{I_t - 1}\mathbf{T}_{c,t}^T\mathbf{T}_{c,t}\right)^{\dagger}\boldsymbol{t}_{new,c,t} \tag{19}$$

$$T_{new,i}^2 = \boldsymbol{t}_{new,i,t}^T\left(\frac{1}{I_t - 1}\mathbf{T}_{i,t}^T\mathbf{T}_{i,t}\right)^{\dagger}\boldsymbol{t}_{new,i,t} \tag{20}$$

Given the confidence level α, the control limit of the Hotelling's T^2 statistics is easy to defined, referring to [33].

The squared prediction error (SPE) statistic of individual subspace reflects the influence of the noise in residual subspace

$$SPE = \left\|\boldsymbol{e}_{new,i,t}\right\|^2 \tag{21}$$

When the calculated statistics are less than the corresponding control limits, it indicates a normal condition; otherwise, there is an anomaly case. Fig. 4 shows the process of offline modeling and online monitoring in this paper.

TABLE I: Description of the fault in numerical example

Fault No.	Fault type	Fault description
Fault 1	Slope drift fault	1^{st} observation variable change by $0.005\,(i-500)$
Fault 2	Step fault	2^{nd} observation variable step change by 3.
Fault 3	Multiplicative fault	The element in the third column of the fourth row of the projection matrix changes from 0.2134 to 0.5134.
Fault 4	Slope drift fault	4^{st} observation variable change by $0.005\,(i-500)$
Fault 5	Step fault	3^{nd} observation variable step change by 2.
Fault 6	Multiplicative fault	The element in the first column of the fifth row of the projection matrix changes from 0.5678 to -0.8423.

V. CASE STUDY

A numerical case and an actual hydrocracking process are applied in this section to demonstrate the effectiveness of the proposed method.

A. Numerical case

A five-variable numerical example can be constructed by the following equation,

$$\begin{bmatrix} \boldsymbol{x}_{s,1} \\ \boldsymbol{x}_{s,2} \\ \boldsymbol{x}_{s,3} \\ \boldsymbol{x}_{s,4} \\ \boldsymbol{x}_{s,5} \end{bmatrix} = \begin{bmatrix} 0.3723 & -0.1893 & 0.4790 \\ 0.9080 & 0.2954 & -0.6489 \\ 0.5842 & 0.3490 & 0.6745 \\ 0.7328 & -0.4389 & 0.2134 \\ 0.5678 & -0.1321 & 0.7882 \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_{s,1} \\ \boldsymbol{s}_{s,2} \\ \boldsymbol{s}_{s,3} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_{s,1} \\ \boldsymbol{e}_{s,2} \\ \boldsymbol{e}_{s,3} \\ \boldsymbol{e}_{s,4} \\ \boldsymbol{e}_{s,5} \end{bmatrix} \tag{22}$$

where we assume each variable $\boldsymbol{x}_{s,i}$ is sampled from a mixture of three Gaussian distributions $\boldsymbol{s}_{s,1} \sim \mathcal{N}(0,1)$, $\boldsymbol{s}_{s,2} \sim \mathcal{N}(1,1)$, and $\boldsymbol{s}_{s,3} \sim \mathcal{N}(2,1)$ plus a white noise $\boldsymbol{e}_{s,i}$ with mean of zero and standard deviations of 0.01. Then the first source domain data $\boldsymbol{x}_s$ is exactly collected with Eq. (22). The second source domain and target domain data are created via merely changing the generation of second variable, respectively, while keeping the remaining ones. That is,

$$\boldsymbol{x}_{s,2} = -0.9080\boldsymbol{s}_{s,1} + 0.2954\boldsymbol{s}_{s,2} - 0.6489\boldsymbol{s}_{s,3} + \boldsymbol{e}_{s,2} \tag{23}$$

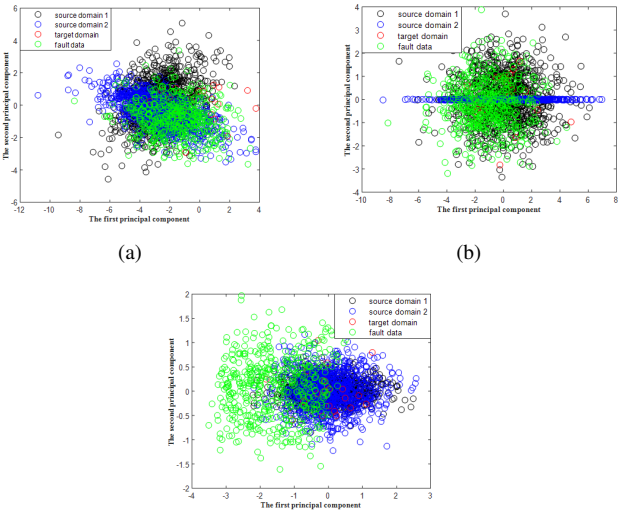
$$\boldsymbol{x}_{t,2} = -0.9080\boldsymbol{s}_{s,1} + 0.2954\boldsymbol{s}_{s,2} + 0.6489\boldsymbol{s}_{s,3} + \boldsymbol{e}_{t,2} \tag{24}$$

Each of the two source domains is with a total of 1000 samples. The target domain is under the condition of few samples, only 20 samples. The common components of historical multi-working conditions are constructed through the data of the source domain, which can be transferred to the few samples working conditions of the target domain. For PCA, sparse PCA [34] and SPCA-SI [35], the two principal components are retained. The value of the common principal subspace of CnI-PCA [22], CPD [21] and the proposed method is also set to two for fair comparison, the coefficient matrix λ is 0.2 by trial and error. In every case, the confidence level of 5% was used to calculate the control limit.

To test the efficiency of the proposed method, six faults, including three types, are introduced under target domain

TABLE II: Monitoring results of numerical example

Fault No.	MsTL method			SPCA-SI		CPD-based method			CnI-PCA			Sparse PCA		PCA	
	T_c^2	T_i^2	SPE_i	T^2	SPE	T_c^2	T_i^2	SPE_i	T_c^2	T_i^2	SPE_i	T^2	SPE	T^2	SPE
Normal	2.0	0.8	1.8	4.6	8.0	1.8	1.4	3.6	3.8	4.6	1.2	4.0	1.2	1.8	1.8
1	4.6	**53.4**	**89.0**	3.6	79.2	2.8	29.0	63.2	5.4	1.4	73.8	9.2	56.8	5.0	57.2
2	25.6	1.0	**100.0**	2.8	54.2	0.8	11.8	81.2	6.8	2.0	46.2	23.2	52.6	25.0	46.4
3	21.4	0.4	**88.4**	76.0	82.2	2.4	18.0	55.8	39.4	68.2	73.0	19.8	26.2	25.2	22.4
4	17.6	24.0	86.2	16.8	65.0	6.6	41.8	47.6	5.8	7.0	65.4	24.8	64.0	17.6	63.4
5	13.6	79.8	**97.4**	2.6	74.4	14.4	24.0	86.2	3.0	2.2	59.2	15.2	79.8	4.0	78.8
6	10.4	20.0	**82.6**	10.4	52.0	4.4	12.0	62.2	5.6	6.4	54.4	9.2	50.0	10.0	49.6



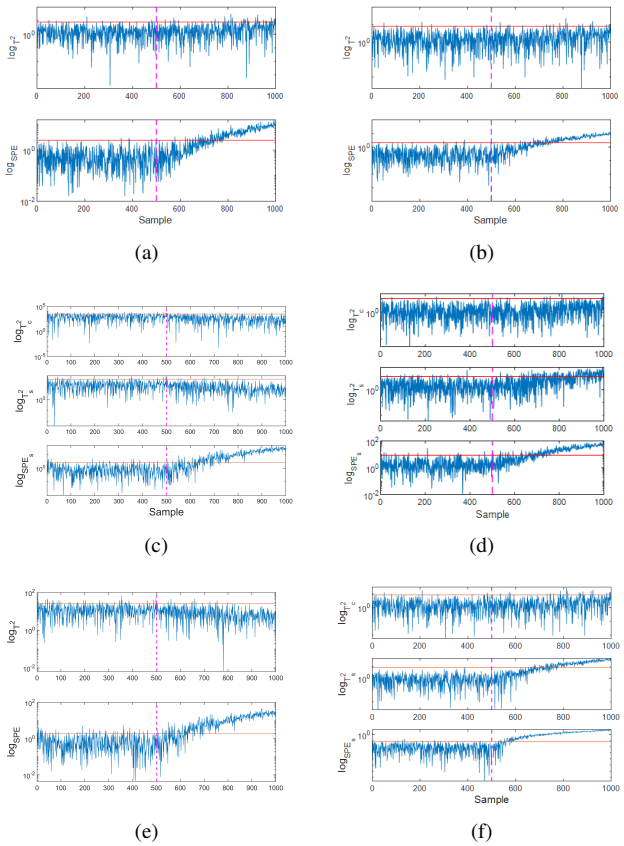
(a) (b) (c)

Fig. 5: (a)Raw data; (b)Common subspace; (c)Individual subspace.

working conditions, as indicated in Table I. The test data are identically distributed with target mode, which has 1000 samples, and faults occur in 500^{th} sample point.

Moreover, the traditional process monitoring methods including PCA and sparse PCA are built using the same 20 samples in target domain only. The multimode process monitoring methods including CnI-PCA, CPD-based method and SPCA-SI are built using 20 samples in each source domain and 20 samples in target domain. The performance of four methods shown in Table II. In Fig. 5, the data distribution of source domain 1, source domain 2, target domain and fault 1 are shown, with original data subspace, common and individual subspace being extracted by the proposed method. It can be seen from Fig. 5(b) that the data distribution of the target domain is close to that of fault 1 in the common subspace, while in the individual subspace Fig. 5(c), the data distribution of the target domain and fault 1 can be clearly distinguished, this means that fault 1 is better monitored in individual subspace. Fig. 6 shows the monitoring performance with the fault 1. It can be seen that in fault 1, the MsTL method achieves a better monitoring effect on the individual subspace T_i^2 and SPE_i, which is consistent with the previous analysis.

From Table II, the proposed MsTL method has the best monitoring effectiveness. The proposed method obtains the



(a) (b) (c) (d) (e) (f)

Fig. 6: (a)PCA; (b)Sparse PCA; (c)CnI-PCA; (d)CPD-based method; (e)SPCA-SI; (f)MsTL method.

TABLE III: Mode description of hydrocracking process

Mode No.	Mode description
Mode 1	Variation in the characteristics of crude oil
Mode 2	Feed variation
Mode 3	Diesel catalytic conversion

common subspace under the new working condition by fusing the prior knowledge of the common components of the historical multiple working conditions. When the training data set is small, the monitoring effectiveness decreases because the PCA and sparse PCA models cannot fully utilize the data in the source domain, and it is difficult to build an accurate model only based on the target domain data of few samples. Similarly, the CnI-PCA, CPD-based method and SPCA-SI model in few samples, and their modeling abilities are affected by the data samples, which cannot reflect the common and individual

TABLE IV: Monitoring results of hydrocracking process

Fault No.	MsTL method			SPCA-SI		CPD-based method			CnI-PCA			Sparse PCA		PCA	
	T_c^2	T_i^2	SPE_i	T^2	SPE	T_c^2	T_i^2	SPE_i	T_c^2	T_i^2	SPE_i	T^2	SPE	T^2	SPE
Normal	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
2	66.7	**100.0**	45.3	79.2	56.9	44.6	45.9	71.6	83.3	**86.1**	79.2	69.3	45.3	65.3	46.7
3	**98.7**	**100.0**	10.7	57.3	16	0.0	9.3	**72.0**	66.7	1.3	21.3	56.0	14.0	30.7	6.8

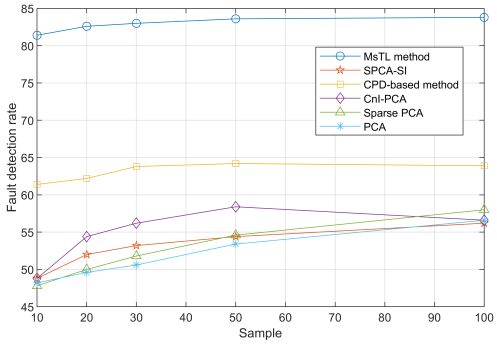


Fig. 7: The sensitivity analysis for the number of samples.

feature of the multimode domain, so their monitoring results are worse than the methods proposed in this paper. At the same time, the monitoring effect of sparse PCA is better than that of PCA, because sparse PCA modifies the principal components with more zero weightings, which makes the model more suitable for small sample data modeling.

In fault 6, the sensitivity analysis is performed on the number of samples from the target domain, as shown in Fig. 7. Let the number of samples change from 10 to 100, the fault detection rates are improved for all methods. From the perspective of fault detection performance saturation, both PCA and sparse PCA show significant improvements in fault detection rates as the number of samples increases. Because their modeling fully depends on the size of the training set of the target domain, the fault detection performance is still not saturated until the sample number is 100. Compared with PCA and sparse PCA, the CnI-PCA and the CPD-based method utilize training data not only from the target domain, but also from the source domain. The prior knowledge of the source domain is extracted by decomposing the data space of the source domain and the target domain into common subspace and individual subspace, so the fault detection performance reaches saturation when the number of target domain samples is 50. Unlike the CnI-PCA and the CPD-based method, the successive historical modes of the source domain are learned in a sequential fashion by the SPCA-SI, whose fault detection performance reaches saturation when the sample is 50. The proposed method introduces transfer learning to more fully fuse the historical data of the source domain to model the new data of the target domain, so the fault detection performance reaches saturation when the sample is 30. In summary, compared with other comparison methods, the proposed approach exhibits the best modeling ability and monitoring performance on the target domain dataset with few samples. The sensitivity analysis for the number of samples verified the effectiveness of the proposed method.

B. Hydrocracking process

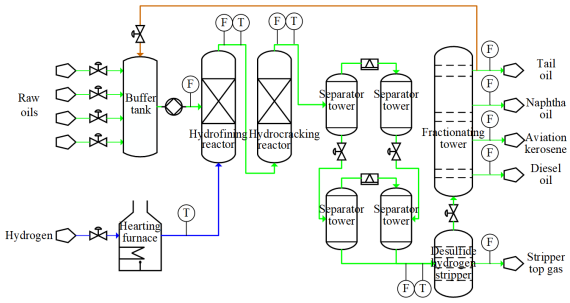


Fig. 8: An industrial hydrocracking process technology flow diagram.

In petrochemical production, hydrocracking reaction is a very representative process. The hydrocracking process used in this paper is carried out from an actual chemical plant. Fig. 8 shows the hydrocracking process flow diagram. Hydrocracking is a petroleum secondary processing method. It mostly turns substandard crude oil into alternative clean and efficient petroleum fuels using chemical processes (hydrofining, hydrocracking) and physical reactions (gas-liquid separation, product fractionation). The hydrocracking and hydrofining process is used to remove the non-hydrocarbon compounds in the feedstock. Hydrocracking is mainly to convert large molecules into small molecules through cracking reaction, so as to further reduce the content of hydrogen, sulfur and nitrogen in the feedstock, and obtain products after catalytic reforming, including heavy naphtha, diesel or jet fuel. After the hydrocracking reaction, the reaction effluent is separated by high- and low-pressure separator. According to the principle of different boiling points of different components, hydrocracking products are fractionated and separated to obtain light liquefied gas, heavy naphtha, kerosene and diesel oil. At the same time, due to the different feed requirements from different processes and the production of different products, hydrocracking needs to change the production process frequently to meet the needs of production of different products, which leads to hydrocracking is a production process with multiple working conditions.

In this study, 188 process variables for the hydrocracking process were chosen. The process data is gathered from industrial historians at 5-minute intervals. Table III shows the selection of three operating modes: variation in the characteristics of crude oil, feed fluctuation, and diesel catalytic conversion. In the test set, there are 108, 120, and 92 samples in working conditions 1, 2, and 3, respectively, and the fault

occurs at the 38th, 48th, and 17th sample points respectively. There are 181 sample points in each source domain in the training set, but only 30 sample points in the target domain, and the target domain is in a few samples environment. To evaluate the effectiveness of the proposed monitoring method, the monitoring model is built using PCA, sparse PCA, CnI-PCA, CPD-based method and SPCA-SI for comparison. The PCA and sparse PCA are built using the same 30 samples in target domain only. The CnI-PCA, CPD-based method and SPCA-SI are built using 30 samples in each source domain and 30 samples in target domain.

For PCA, sparse PCA and SPCA-SI, the number of principle components is set to 15 to cover 85% of the sum of eigenvalues of the covariance matrix. For CnI-PCA and CPD-basd method, the dimensionality of common subspace and individual subspace is set to 15, and the value of the common principal subspace of the proposed method is also set to 15 for fair comparison, the coefficient matrix λ is 0.1 by trial and error. In every case, the confidence level of 5% was used to calculate the control limit.

Fig. 9 depicts the data distribution of the source domains 1, 2, and 3 as well as fault 3, with original data subspace, common and individual subspace are extracted by proposed method. And from Fig. 9, the fault data and target domain data are well differentiated in both the common subspace and the individual subspace, which is consistent with the monitoring effect of the proposed method demonstrated in Fig. 10(f). Table IV's performance of the six approaches reveals that the proposed method outperforms them all. The proposed method obtains the common subspace of the target domain by incorporating the common components in the source domain as a priori knowledge. Therefore, it still achieves good monitoring performance on the target domain dataset with few samples, which indicates that the common subspace of the target domain effectively integrates the prior knowledge of the common subspace of the source domain. For PCA and sparse PCA, the monitoring performance is poor in the case of few samples.Similarly, the CnI-PCA and CPD-based method model in few samples, and their modeling abilities are affected by few data samples. Although SPAC-SI is a few-sample method, it may forget the knowledge learned from the historical modes, so it is no more effective than the proposed method. As for the sparse PCA, if there are much more variables than samples, the sparse method gets modified PCs with some possibly zero weightings, and the monitoring effectiveness is better than traditional PCA.

VI. CONCLUSIONS

In this paper, a MsTL method is proposed to transfer the common components of historical multimode as prior knowledge to the few-samples mode in the target domain. It solves the problem that the modeling ability declines due to insufficient data at the early stage of the new domain. The optimal transformation of matching new mode features with common features is achieved without changing the internal association of new modalities. We verify the effectiveness of the MsTL method through a numerical example and a real-world hydrocracking case. In future work, we will further

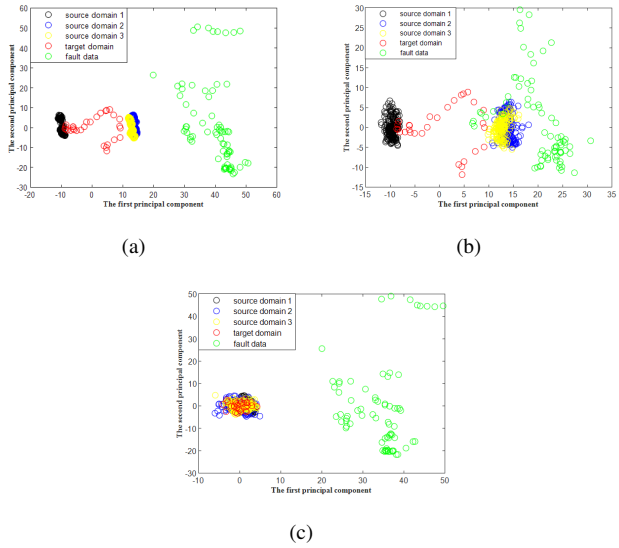


Fig. 9: (a)Raw data; (b)Common subspace; (c)Individual subspace.

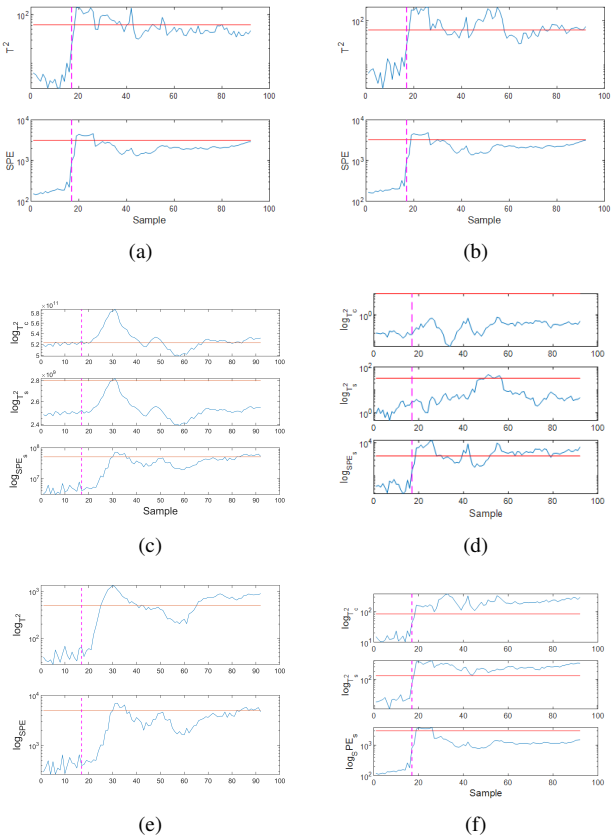


Fig. 10: (a)PCA; (b)Sparse PCA; (c)CnI-PCA; (d)CPD-based method; (e)SPCA-SI; (f)MsTL method.

exploit this problem under nonlinear situations at the data level and complex dynamic characteristics at the process level.

REFERENCES

[1] K. Wang, J. Chen, Z. Song, Y. Wang, and C. Yang, "Deep neural network-embedded stochastic nonlinear state-space models and their applications to process monitoring," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 7682–7694, 2021.

[2] K. Huang, H. Wen, C. Zhou, C. Yang, and W. Gui, "Transfer dictionary learning method for cross-domain multimode process monitoring and fault isolation," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 11, pp. 8713–8724, 2020.

[3] S. Yin, G. Wang, and H. Gao, "Data-driven process monitoring based on modified orthogonal projections to latent structures," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 4, pp. 1480–1487, 2016.

[4] Q. Jiang, X. Yan, and B. Huang, "Performance-driven distributed pca process monitoring based on fault-relevant variable selection and bayesian inference," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 1, pp. 377–386, 2015.

[5] M.-F. Harkat, M. Mansouri, M. N. Nounou, and H. N. Nounou, "Fault detection of uncertain chemical processes using interval partial least squares-based generalized likelihood ratio test," *Information Sciences*, vol. 490, pp. 265–284, 2019.

[6] R. Zhai, J. Zeng, and Z. Ge, "Structured principal component analysis model with variable correlation constraint," *IEEE Transactions on Control Systems Technology*, vol. 30, no. 2, pp. 558–569, 2022.

[7] Y. Qin, Y. Yan, H. Ji, and Y. Wang, "Recursive correlative statistical analysis method with sliding windows for incipient fault detection," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 4, pp. 4185–4194, 2022.

[8] L. Luo, W. Wang, S. Bao, X. Peng, and Y. Peng, "Robust and sparse canonical correlation analysis for fault detection and diagnosis using training data with outliers," *Expert Systems with Applications*, vol. 236, p. 121434, 2024.

[9] Y. Lyu, L. Zhou, Y. Cong, H. Zheng, and Z. Song, "Multirate mixture probability principal component analysis for process monitoring in multimode processes," *IEEE Transactions on Automation Science and Engineering*, pp. 1–12, 2023.

[10] Z. Zhang, P. Chen, C. Xing, B. Liu, R. Wang, L. Li, X. Chen, and E. Zio, "A data augmentation boosted dual informer framework for the performance degradation prediction of aero-engines," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 12 018–12 030, 2023.

[11] A. Mehrotra and A. Dukkipati, "Generative adversarial residual pairwise networks for one shot learning," *CoRR*, vol. abs/1703.08033, 2017. [Online]. Available: http://arxiv.org/abs/1703.08033

[12] W. Li, J. Chen, J. Cao, C. Ma, J. Wang, X. Cui, and P. Chen, "Eid-gan: Generative adversarial nets for extremely imbalanced data augmentation," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 3, pp. 3208–3218, 2023.

[13] S. Bartunov and D. Vetrov, "Few-shot generative modelling with generative matching networks," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Storkey and F. Perez-Cruz, Eds., vol. 84. PMLR, 09–11 Apr 2018, pp. 670–678. [Online]. Available: https://proceedings.mlr.press/v84/bartunov18a.html

[14] Q. Sun, Y. Liu, Z. Chen, T.-S. Chua, and B. Schiele, "Meta-transfer learning through hard tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1443–1456, 2022.

[15] Y. Hu, R. Liu, X. Li, D. Chen, and Q. Hu, "Task-sequencing meta learning for intelligent few-shot fault diagnosis with limited data," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3894–3904, 2022.

[16] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *International Conference on Machine Learning*, 2007.

[17] T. Zhang, J. Chen, S. Liu, and Z. Liu, "Domain discrepancy-guided contrastive feature learning for few-shot industrial fault diagnosis under variable working conditions," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 10, pp. 10 277–10 287, 2023.

[18] H. Wang, J. Wang, Y. Zhao, Q. Liu, M. Liu, and W. Shen, "Few-shot learning for fault diagnosis with a dual graph neural network," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1559–1568, 2023.

[19] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1182–1194, 2018.

[20] F. Liu, G. Zhang, and J. Lu, "Heterogeneous domain adaptation: An unsupervised approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5588–5602, 2020.

[21] K. Zhang, K. Peng, S. Zhao, and F. Wang, "A novel feature extraction-based process monitoring method for multimode processes with common features and its applications to a rolling process," *IEEE Transactions on Industrial Informatics*, vol. PP, no. 99, pp. 1–1, 2020.

[22] K. Zhang, K. Peng, and J. Dong, "A common and individual feature extraction-based multimode process monitoring method with application to the finishing mill process," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4841–4850, 2018.

[23] C. Zhao, F. Gao, D. Niu, and F. Wang, "A two-step basis vector extraction strategy for multiset variable correlation analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 147–154, 2011.

[24] C. Zhao and F. Gao, "Two-step multiset regression analysis (msra) algorithm," *Industrial & Engineering Chemistry Research*, vol. 51, no. 3, pp. 1337–1354, 2012. [Online]. Available: https://doi.org/10.1021/ie201608f

[25] J. Liu, J. Hou, and J. Chen, "Dual-layer feature extraction based soft sensor methods and applications to industrial polyethylene processes," *Computers & Chemical Engineering*, vol. 154, p. 107469, 2021.

[26] R. Zdunek, K. Fonał, and A. Wołczowski, "Linked cp tensor decomposition algorithms for shared and individual feature extraction," *Signal Processing: Image Communication*, vol. 73, pp. 37–52, 2019.

[27] H. Cheng, Y. Liu, D. Huang, Y. Pan, and Q. Wang, "Adaptive transfer learning of cross-spatiotemporal canonical correlation analysis for plant-wide process monitoring," *Industrial & Engineering Chemistry Research*, vol. 59, no. 49, pp. 21 602–21 614, 2020.

[28] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[29] J. Zhu and F. Gao, "Similar batch process monitoring with orthogonal subspace alignment," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 10, pp. 8173–8183, 2018.

[30] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[31] T. Bakhtiar, "Orthogonal procrustes analysis: Its transformation arrangement and minimal distance," *International Journal of Applied Mathematics and Statistics*, vol. 20, no. M11, pp. 16–24, 2010.

[32] C. Yang, H. Liang, K. Huang, Y. Li, and W. Gui, "A robust transfer dictionary learning algorithm for industrial process monitoring," *Engineering*, vol. 7, no. 9, pp. 1262–1273, 2021.

[33] K. Wang, W. Zhou, Y. Mo, X. Yuan, Y. Wang, and C. Yang, "New mode cold start monitoring in industrial processes: A solution of spatial–temporal feature transfer," *Knowledge-Based Systems*, vol. 248, p. 108851, 2022.

[34] S. Gajjar, M. Kulahci, and A. Palazoglu, "Real-time fault detection and diagnosis using sparse principal component analysis," *Journal of Process Control*, vol. 67, pp. 112–128, 2018.

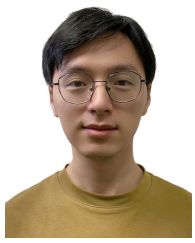
[35] J. Zhang, D. Zhou, and M. Chen, "Self-learning sparse pca for multimode process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 29–39, 2023.

Kai Wang received the B.Eng. and Ph.D. degrees from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2014 and 2019, respectively. He was a visiting scholar with the Department of Chemical and Biological Engineering at the University of British Columbia. He is currently a Professor with the School of Automation, Central south University. He specializes in industrial data analytics, process health management and machine learning.

Xiang Lei received the B.Eng. degree in automation from the School of Automatic and Electronic Information, Xiangtan University, Xiangtan, China, in 2022. He is currently pursuing the master's degree with the School of Automation, Central South University, Changsha, China. His research interests include process monitoring, soft sensor modeling, process data mining, and deep learning.

Wenxuan Zhou received the B.Eng. And M.Eng. degrees from the School of Automation, Central South University, Changsha, China, in 2020 and 2023, respectively. His research interests include industrial data analytics and process monitoring.

Saige Chen received the B.Eng. and M.Eng degrees from Northeast Forestry University, Harbin, China, in 2020 and 2022, respectively. Since 2022, he is studying for an Ph.D. degree with the School of Automation, Central South University, Changsha, China. His current research interests include advanced process control, reinforcement learning control, and data rectification.



Jing Li is working as a research scientist at Centre for Frontier AI Research (CFAR) in Agency for Science, Technology and Research (A*STAR), Singapore. He obtained his Ph.D. degree in April 2023 from University of Technology Sydney (UTS), Australia. Beofre that, he received the B.Eng and M.Eng degrees in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2015 and 2018, respectively. He has published papers in top journals and conferences, such as JMLR, IEEE TPAMI, IEEE TIP, IEEE TIFS, IEEE TKDE, IEEE TNNLS, IJCAI, and IEEE ICASSP. His research mainly focus on trustworthy machine learning.