Principal Component-Based Semi-Supervised Extreme Learning Machine for Soft Sensing

Xudong Shi, *Graduate Student Member*, Qi Kang, *Senior Member, IEEE*, Hanqiu Bao, *Graduate Student Member*, Wangya Huang, and Jing An, *Member, IEEE*

***Abstract*— Soft sensing technique has been extensively used to predict key quality variables in industrial systems. However, due to the difficulty of quality variable acquisition, only limited labeled data samples are available, and a large number of unlabeled ones are discarded. This raises a big challenge to build a high-quality soft sensor model. In order to furthest exploit information contained in both the labeled and unlabeled data, this paper proposes a principal component-based semi-supervised extreme learning machine (referred to as PCSELM) model. Through this model, extracting latent features and learning nonlinear input-output relationship can be simultaneously performed. In this way, unlabeled samples are utilized efficiently for feature representation and model accuracy improvement. Moreover, mixed regularizations are employed to work in conjunction with the PCSELM to obtain high generality and flexibility. We also derive an efficient parameter learning algorithm with theoretically guaranteed convergence. Comprehensive experiments are conducted via an industrial process. Comparison results illustrate that the proposed PCSELM outperforms other representative semi-supervised algorithms.**

***Note to Practitioners*—Industrial processes in general incorporate unlabeled samples which are ubiquitous in real world applications. The focus of this paper is to develop a semi-supervised soft sensor model (PCSELM) that is capable to learn the nonlinear features and regression relationship efficiently with both the labeled and unlabeled samples. The proposed model can automatically implement the feature representation and the input-output relationship description. In addition, we introduce mixed norms for the model objective function to improve the final prediction performance and generalization. A feasible model optimization technique with proved convergence is also derived. Experimental results based on a real industrial dataset manifest that PCSELM achieves better prediction accuracy than its peers.**

***Index Terms*— Soft sensing, semi-supervised learning, extreme learning machine, principal component, regularization.**

Manuscript received 25 April 2023; accepted 21 June 2023. This article was recommended for publication by Associate Editor Z. Kong and Editor L. Zhang upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 51775385, in part by the Natural Science Foundation of Shanghai under Grant 23ZR1466000, in part by the Shanghai Industrial Collaborative Science and Technology Innovation Project under Grant 2021-cyxt2-kj10, in part by the Innovation Program of Shanghai Municipal Education Commission under Grant 202101070007E00098, in part by the Fundamental Research Funds for the Central Universities under Grant 2023-4-YB-07, and in part by the Research Foundation of the Shanghai Institute of Technology under Grant KJFZ2023-10. *(Corresponding author: Qi Kang.)*

Xudong Shi, Qi Kang, and Hanqiu Bao are with the Department of Control Science and Engineering and the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, China (e-mail: xdshi@tongji.edu.cn; qkang@tongji.edu.cn; 1910637@tongji.edu.cn).

Wangya Huang is with the Department of Control Science and Engineering, Tongji University, Shanghai 201804, China, and also with the Silicon Steel Business Unit, Baoshan Iron & Steel Company Ltd., Shanghai 201900, China (e-mail: huangwy@baosteel.com).

Jing An is with the School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai 201418, China (e-mail: anjing@sit.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2023.3290352.

Digital Object Identifier 10.1109/TASE.2023.3290352

I. Introduction

FOR the high-level quality products, economic profits enhancement and process safety maintenance of industrial processes, it is necessary to implement the valid process monitoring [1], [2], control, and optimization [3], [4] in the area of industrial internet of things [5], [6], [7], [8], [9]. However, technical and economical limitations such as the low reliability of sensing devices and expensive time and human resources impose constrains on the hard-ware sensors to provide the real time measuring of quality variables [10], [11], [12], [13], [14], which could lead to production quality deteriorations, energy consumptions, or even process safety risks [15], [16], [17], [18], [19], [20]. In practice, the precise online measuring technique for quality variables is desired.

With the advancements of data acquisition and industrial intelligence, data-driven soft sensors constructing the predictive mathematical models between the hard-to-measure quality variables (i.e., output variables) and easy-to-measure process variables (i.e., input variables), have a provision of economical alternatives to those expensive physical sensors [21], [22]. Soft sensors have been widely studied and applied for quality prediction and have received increased attentions in both the academia and the industry area, in view of their convenience of modeling, no time delay and low cost. Frequently used methods have been applied for soft sensor modeling, such as principal component analysis (PCA) [23], [24], partial least squares (PLS) [25], [26], canonical correlation analysis (CCA) [27], [28], extreme learning machine (ELM) [29], [30], [31], [32], etc.

Owing to the merits of nonlinear fitting capability and extremely fast learning speed, ELM has been widely adopted for soft sensing of complex industrial processes [33], [34], [35], [36]. Notwithstanding the accurate prediction performance achieved by the aforementioned ELM-based soft sensors, there still exist some limitations associated with them. Firstly, randomly configured parameters could lead to strong-coupled and redundant features (i.e., only a few

essential generated features dominate data variations independently), which often cause the over-fitting issue and greatly affects the performance of the ELM. Therefore, it is desired and necessary to perform the feature representation or dimension reduction algorithms for ELM to exploit a latent data structure [37], [38], [39], [40], [41], [42], [43] that truly represent the nature of the dataset.

Moreover, the scenario of limited labeled samples may further prevent the ELM from achieving satisfactory prediction performances due to unreliable parameter estimations. Despite the labeled samples are limited, there are numerous unlabeled samples. Therefore, how to effectively utilize the unlabeled samples with labeled samples together determines the accuracy of soft sensor [44], [45]. In machine learning field, the paradigm which trains a model with labeled data is termed as the supervised learning, while the paradigm that trains a model with both of labeled and unlabeled data is terms as the semi-supervised learning [46], [47]. Since those quality variables to be estimated is quite hard to collect, the available labeled samples are rare. The semi-supervision is a nature in vast soft sensor applications.

In order to address the problem of soft sensing nonlinear industrial processes with limited labeled samples, we propose a method, named principal component-based semi-supervised extreme learning machine (PCSELM). Our main idea is to conduct the feature extraction and semi-supervised learning procedure in ELM framework by embedding three learnable matrices. It should be noted that all the supervised, semi-supervised, and unsupervised ELMs can be included into an unified framework which consists of two stages: 1) random feature mapping and 2) output weights optimization. The hidden layer neurons are randomly generated and fixed during the training process. This key property enables ELM for fast nonlinear feature representation and differs from that of many existing neural models. The second stage is to solve the output weights. This is where the main difference among different ELM variants lies. To this end, the PCSELM improve the training process of ELM, which achieves combined objectives: feature learning and semi-supervised regression.

Specifically, we design three matrices to formulate the proposed model, which are the projection matrix for feature extraction, data recovery matrix for data reconstruction, and regression matrix for quality prediction. By doing so, the latent data structure extraction and input-output regression issues can be both taken good care of. In addition, the PCSELM incorporates both of the scarce labeled and abundant unlabeled data samples, and assign proper matrix norm-based regularizations for generality and regression power enhancement. In detail, a nuclear norm-based regularization and a $L_{2,p}$ norm-based one are attached to the projection and regression matrix, respectively. The former can attain the extracted features with a low-rank data structure and the latter can greatly enhances generality of output results. Meanwhile the data recover matrix is over an orthonormal constrain to reveal some underlying affinity properties. For optimal parameters learning, a new parameter learning algorithm is derived, in which gradient descend and equivalent optimization problem conversion are incorporated. As a result, the proposed PCSELM can be optimized efficiently with theoretical guarantee.

The main novel contributions of this work can be summarized as follows:

1) A novel principal component based semi-supervised extreme learning machine (PCSELM) framework is proposed for industrial soft sensing. The learning of low-dimensional feature representation and disentangling of nonlinear regression relationship is naturally unified through the proposed framework, which enables unlabeled samples to enhance the model reliability and prediction performance.

2) To further improve the generality of the PCSELM, mixed regularizations are designed to offers a new extreme learning paradigm for semi-supervised regression which is endowed with high flexibility and expressive ability. A parameter optimization algorithm for the PCSELM under the framework of block-wise coordinate descent is also derived, in which both the labeled and unlabeled data are used and the parameters convergence are theoretically guaranteed.

The remainder of this article is organized as follows. Section II briefly reviews the related works and the basic ELM model. Section III presents the proposed PCSELM in detail including the model construction and parameter learning, and how to perform soft sensing. An industrial case study is carried out in Section IV. Eventually, some conclusions are drawn in Section V.

II. Backgrounds

A. Related Works

In this section, we briefly review the researches focusing on ELM and semi-supervised learning under the background of soft sensing. ELM is originally proposed by Huang et al. [32] as a learning scheme for single layer feedforward neuron networks. It achieves high training speed since only output weights are optimized. Despite the simple configuration, ELM has the universal nonlinear approximation and good regression capability. As a result, ELM has been widely applied for soft sensor modeling of nonlinear industrial processes. For instance, Geng et al. [30] proposed an self-organizing ELM via the biological neuron-glia interaction principle; Zhang et al. [33] constructed a new similarity measure criterion that considering variable and sample information to construct a double-level locally weighted ELM. Ouyang et al. [34] proposed an advanced deep ELM network algorithm for measuring nitrogen oxides (NOx) concentrations in vehicle exhaust.

The other topic related to this work is semi-supervised learning. Pseudo labeling is the most Intuitive and representative semi-supervised learning diagram, which tries to provide reliable pseudo-labels for the unlabeled data samples. For example, Ge et al. [48] proposed an self-training-based PLS model; Li et al. [49] proposed semi-supervised multiple-output models via the PLS and RVM-based co-training and tri-training diagram, Feng et al. [50] proposed an adversarial smoothing tri-regression model, Sun et al. [51] employed the kernel density estimation and Bayesian learning algorithm for semi-supervised soft sensing. These works pay attention to providing accurate pseudo labels, therefore their performances are solely depends on the estimation of the missing labels. Capturing the correlation between the collected inputs-outputs pairs is helpful to exploiting useful information

of the unlabeled data, so as to improve the final prediction performance. Shao et al. [52] proposed semi-supervised Student's t mixture model for non-Gaussian process soft sensing, in which the unlabeled data samples are used for estimating the reliable and correct probability distribution functions. Ren et al. [53] combined kernel risk-sensitive loss and hyper-graph regularized to develop a new robust extreme learning machine. All of those works have further helped soft sensors to meet the requirements for practical industrial tasks.

B. Preliminary: ELM

As a single-hidden-layer feedforward neuron network, extreme learning machine (ELM) is comprised of one input layer, one output layer representing output, and one hidden layer connecting the input and output layers.

Given an ELM with D input nodes, L hidden nodes and M output nodes, the input vector $\boldsymbol{z} \in \mathbb{R}^D$ is mapped into the feature space, which is given by

$$\boldsymbol{x}(\boldsymbol{z}) = \left[g\left(\boldsymbol{u}_1^\top \boldsymbol{z} + v_1\right), \cdots, g\left(\boldsymbol{u}_L^\top \boldsymbol{z} + v_L\right)\right] \in \mathbb{R}^L \tag{1}$$

where $\{\boldsymbol{u}_l, v_l\}_{l=1}^L \subset \mathbb{R}^D \times \mathbb{R}$ mean the input weights and biases which are randomly generated, and $g(\cdot)$ denotes the activation function.

Let $\left\{\boldsymbol{z}_n, \boldsymbol{y}_n\right\}_{n=1}^N \subset \mathbb{R}^D \times \mathbb{R}^M$ be a training dataset, where $\boldsymbol{z}_n$ is the nth D-dimensional input vector and $\boldsymbol{y}_n$ is the corresponding M-dimensional output vector. The objective function of an ELM is given as follows

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\left\|\boldsymbol{Y} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_F^2 \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{(L+1)\times M}$ means output regression coefficients, $\boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{z}_1); \boldsymbol{x}(\boldsymbol{z}_2); \cdots; \boldsymbol{x}(\boldsymbol{z}_n)]^\top$, $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{1}_N]$, $\boldsymbol{1}_N = [1, 1, \cdots, 1] \in \mathbb{R}^N$, and λ are the regularization parameters.

Solving (2) results in

$$\boldsymbol{\beta} = \left(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}} + \lambda \boldsymbol{I}_{L+1}\right)^{-1} \widetilde{\boldsymbol{X}}\boldsymbol{Y} \tag{3}$$

where $\boldsymbol{I}_{L+1}$ represents the unit matrices with order of $L+1$.

III. Methodology

The core to successful soft sensor modeling is to discover authentic regression relationship between the process variables and quality variables. However, the classical ELM is not competent, because its hidden nodes are high-dimensional and strong-coupled, which may lead to the over-fitting and poor-generalization issues, especially when the labeled data samples are scarce. Motivated by the concept of ELM and principal component analysis, this section proposes a regression model named as principal component based semi-supervised ELM (PCSELM). Three matrices including the projection, recovery, and regression ones are designed and learned via an optimization composed of one newly constructed objective function subject to one constraint.

The objective function consists of three parts: (i) J_X represents for the dimension-reduction by projecting the input variables to the latent space to extract the principal components; (ii) J_Y represents for regression by minimizing the regression errors between the projected input variables and output variables, where both labeled and unlabeled data samples are incorporated; (iii) J_P represent for parameters regularization or penalty which is of great importance for alleviating over-fitting. One constraint is implemented to orthogonalize the projection matrix. How to structure and train the PCSELM, as well as how to develop soft sensor using it are detailed in the rest parts of this section.

A. Formulation of the PCSLEM

Given labeled and unlabeled data matrices as $\left\{\boldsymbol{Z}^l, \boldsymbol{Y}\right\} = \left\{\boldsymbol{z}_i, \boldsymbol{y}_i\right\}_{i=1}^{N_l}$ and $\boldsymbol{Z}^u = \left\{\boldsymbol{z}_j\right\}_{j=N_l+1}^{N_l+N_u}$, respectively, where N_l and N_u are numbers of labeled and unlabeled samples. Usually, $N_u \gg N_l$ holds. For simplicity, we denote N as $N_l + N_u$. First, the hidden layer maps the input vector $\boldsymbol{z}_k$ as $\boldsymbol{x}_k$ by (1), where $k = 1, 2, \cdots, N$. Denote $\boldsymbol{B} \in \mathbb{R}^{L\times R}$, $\boldsymbol{A} \in \mathbb{R}^{L\times R}$, and $\boldsymbol{D} \in \mathbb{R}^{R\times M}$ as the projection, recovery, and regression matrices, where R is the number of the principal components.

(i) Specifically, the constructions of J_X is inspired by the classical principal component analysis to minimize the reconstruction errors, which is given by

$$\begin{aligned} J_X &= \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right\|_F^2 \\ \text{s.t.} &\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}_R \end{aligned} \tag{4}$$

where $\boldsymbol{X} = \left[\boldsymbol{X}^l, \boldsymbol{X}^u\right] \in \mathbb{R}^{N\times L}$, $\boldsymbol{X}\boldsymbol{B} \in \mathbb{R}^{N\times R}$ is considered as principal components, $\boldsymbol{I}_R$ represents the unit matrices with order of R. $\boldsymbol{X}^l$ and $\boldsymbol{X}^u$ represent the output of hidden layer with respect to $\boldsymbol{Z}^l$ and $\boldsymbol{Z}^u$, i.e., $\boldsymbol{x}(\boldsymbol{Z}^l)$ and $\boldsymbol{x}(\boldsymbol{Z}^u)$, where $\boldsymbol{x}(\cdot)$ denotes PCSELM's hidden layer as shown in (1).

(ii) To extract regression relationship among input variables and output variables, the constructions of J_Y is through minimizing the prediction errors, which leads to

$$J_Y = \frac{1}{2}\left\|\boldsymbol{C}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{D} - \boldsymbol{1}_N\boldsymbol{b}^\top\right)\right\|_F^2 \tag{5}$$

where $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^{N\times M}$ is the augmented output matrix designed for semi-supervised regression whose first N_l rows equal to $\boldsymbol{Y}$ and the rest equal to $\boldsymbol{0}$, $\boldsymbol{C} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose first N_l diagonal entries are 1 and the rest are 0, $\boldsymbol{b} \in \mathbb{R}^{M\times 1}$ is the bias.

(iii) Two parameter regularizations are expressed as follows

$$J_{P_1} = \frac{1}{2}\|\boldsymbol{B}\|_*, \quad J_{P_2} = \frac{1}{2}\|\boldsymbol{D}\|_{2,p}^p \tag{6}$$

where the former encourages sparsity of singular values, which achieves a low rank property, while the latter ensures row-wise sparsity, which guarantees generalization of prediction performance.

The used matrix norms are explained as follows: $\|\cdot\|_F$, $\|\cdot\|_*$, and $\|\cdot\|_{2,p}$ are the Frobenius, nuclear, and $L_{2,p}$ norms, where $0 < p < 2$ is an adjustable parameter. Concretely, given a matrix $\boldsymbol{\Theta} \in \mathbb{R}^{I\times J}$, its Frobenius, nuclear, and $L_{2,p}$ norms [41], [42], [43] can be calculated as follows

$$\|\boldsymbol{\Theta}\|_F = \left(\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\theta_{ij}\right)^2\right)^{\frac{1}{2}} \tag{7}$$

$$\|\boldsymbol{\Theta}\|_* = Tr\left(\sqrt{\boldsymbol{\Theta}^\top\boldsymbol{\Theta}}\right) = \sum_{i=1}^{r}\sigma_i \tag{8}$$

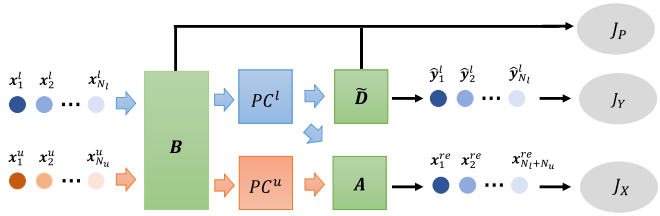


Fig. 1. The architecture of PCSELM.

$$\|\boldsymbol{\Theta}\|_{2,p} = \left(\sum_{i=1}^{I}\left(\sum_{j=1}^{J}(\theta_{ij})^2\right)^{\frac{p}{2}}\right)^{\frac{1}{p}} \tag{9}$$

where $Tr(\cdot)$ denotes the operator for computing the trace of an squared matrix, $r = \min(I, J)$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are $\boldsymbol{\Theta}$'s singular values.

By combining the three parts of objective function ((4), (5), (6)), the PCSELM is formulated as

$$\begin{aligned}\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{D},\boldsymbol{b}} \quad & J = \xi J_X + \xi J_Y + \lambda\left(J_{P_1} + J_{P_2}\right) \\ & = \frac{\xi}{2}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{B}\|_* \\ & + \frac{\xi}{2}\left\|\boldsymbol{C}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{D} - \boldsymbol{1}_N\boldsymbol{b}^\top\right)\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{D}\|_{2,p}^p \\ \text{s.t.} \quad & \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}_R \end{aligned} \tag{10}$$

The overall architecture of the proposed approach is shown in Fig. 1, where PC^l and PC^u means the extracted principal components of the labeled and unlabeled samples, respectively, $\hat{y}^l$ is the prediction for the labeled samples, and $\boldsymbol{x}^{re}$ denotes the reconstruction for the labeled and unlabeled samples.

Note that PCSELM is comprised of two data transformation operations, i.e., the dimension reduction and the regression, which is presented by the two error-based terms. The former is for data reconstruction and the latter is for quality prediction. Moreover, one can find that both the labeled and unlabeled samples take effect in learning model parameters. We thereby expect the extracted principal components are capable to represent the essential information of the process data, also the functional relationship between the input and output can be taken good care of. However, the optimization (10) cannot be solved by common-used mathematical tools. Some deductions should be conducted to solve it, hence an efficient optimization approach is developed next. To solve problem (10), a suitable way is to derive a block-wise coordinate descent-based optimization algorithm, where $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$ and $\boldsymbol{b}$ is optimized iteratively while the other parameters are fixed. Some deductions of the parameter optimization approach are detailed next.

B. Model Optimization

In this subsection, we present an efficient optimization approach for solving (10), which leads to

$$\begin{aligned}\min \quad & J = \frac{\xi}{2}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{E}\|_* \\ & + \frac{\xi}{2}\left\|\boldsymbol{C}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{D} - \boldsymbol{1}_N\boldsymbol{b}^\top\right)\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{D}\|_{2,p}^p \\ \text{s.t.} \quad & \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}_R, \boldsymbol{B} = \boldsymbol{E} \end{aligned} \tag{11}$$

where $\boldsymbol{E} \in \mathbb{R}^{L\times R}$ is an introduced variable to make the objective function separable.

(11) can be solved by optimizing the following augmented Lagrange multiplier (ALM) problem

$$\begin{aligned}&\mathcal{L}_A(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{E}; \boldsymbol{W}) \\ &= \frac{\xi}{2}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{E}\|_* \\ &+ \frac{\xi}{2}\left\|\boldsymbol{C}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{D} - \boldsymbol{1}_N\boldsymbol{b}^\top\right)\right\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{D}\|_{2,p}^p \\ &- Tr\left(\boldsymbol{W}^\top(\boldsymbol{B} - \boldsymbol{E})\right) + \frac{\mu}{2}\|\boldsymbol{B} - \boldsymbol{E}\|_F^2 \end{aligned} \tag{12}$$

where $\mu > 0$ is a penalty parameter, and $\boldsymbol{W} \in \mathbb{R}^{L\times R}$ is an associated Lagrange multiplier. It is worthy to notice that the above expression is under the $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}_R$ constrain. To solve (12), we separate it into the following subproblems.

1) $\boldsymbol{A}$-Subproblem: By fixing the other variables, it leads to

$$\begin{aligned}\boldsymbol{A}^* &= \arg\min_{\boldsymbol{A}^\top\boldsymbol{A}=\boldsymbol{I}}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right\|_F^2 \\ &= \arg\min_{\boldsymbol{A}^\top\boldsymbol{A}=\boldsymbol{I}} Tr\left[\left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right)^\top\left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\right)\right] \\ &= \arg\max_{\boldsymbol{A}^\top\boldsymbol{A}=\boldsymbol{I}} Tr\left(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\right) \end{aligned} \tag{13}$$

$\boldsymbol{A}$-subproblem is a Procrustes rotation problem, and Proposition 1 shows how to solve it.

[Proposition 1]: Let $\boldsymbol{G}, \boldsymbol{N} \in \mathbb{R}^{D\times R}$. Consider the following constrained optimization problem

$$\boldsymbol{G}^* = arg\max_{\boldsymbol{G}^T\boldsymbol{G}=\boldsymbol{I}} Tr\left(\boldsymbol{G}^\top\boldsymbol{N}\right) \tag{14}$$

Suppose the singular value decomposition (SVD) of $\boldsymbol{N}$ results in $\boldsymbol{N} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, then $\boldsymbol{G}^* = \boldsymbol{U}\boldsymbol{V}^\top$.

The proof details can be found in [40].

According to Proposition 1, $\boldsymbol{A}$-subproblem's closed-form solution is

$$\boldsymbol{A}^* = \boldsymbol{U}_1\boldsymbol{V}_1^\top \tag{15}$$

where $\boldsymbol{U}_1$, $\boldsymbol{V}_1$ are calculated via performing SVD on $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}$, i.e., $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top$

2) $\boldsymbol{B}$-Subproblem: While fixing the other variables, it becomes

$$\begin{aligned}\boldsymbol{B}^* = \arg\min_{\boldsymbol{B}} & \frac{\xi}{2}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\right\|_F^2 \\ & + \frac{\xi}{2}\left\|\boldsymbol{C}^{\frac{1}{2}}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{D} + \boldsymbol{1}_N\boldsymbol{b}^\top\right)\right\|_F^2 \\ & - Tr\left(\boldsymbol{W}^\top(\boldsymbol{B} - \boldsymbol{E})\right) + \frac{\mu}{2}\|\boldsymbol{B} - \boldsymbol{E}\|_F^2 \end{aligned} \tag{16}$$

Take the derivative of (16) with respective to $\boldsymbol{B}$, and denote it as $\mathcal{G}(\boldsymbol{B})$, we have

$$\begin{aligned}\mathcal{G}(\boldsymbol{B}) &= \xi\left(-\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A} + \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top\boldsymbol{A}\right) \\ &+ \xi\left(-\boldsymbol{X}^\top\boldsymbol{C}\widetilde{\boldsymbol{Y}}\boldsymbol{D}^\top + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}\boldsymbol{B}\boldsymbol{D}\boldsymbol{D}^\top - \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{1}_N\boldsymbol{b}^\top\boldsymbol{D}^\top\right) \\ &- \boldsymbol{W} + \mu(\boldsymbol{B} - \boldsymbol{E}) \end{aligned} \tag{17}$$

Hence, $\boldsymbol{B}$ can be optimized via a gradient descent manner, i.e.,

$$\boldsymbol{B}^{k+1} = \boldsymbol{B}^k - \eta\mathcal{G}\left(\boldsymbol{B}^k\right) \tag{18}$$

where η denotes learning rate and k indicates the current iteration.

3) $\boldsymbol{D}$ and $\boldsymbol{b}$-Subproblems: For the sake of simplicity, let $\widetilde{\boldsymbol{X}}^l = \left[\boldsymbol{1}_{N_l}, \boldsymbol{X}^l\boldsymbol{B}\right] \in \mathbb{R}^{N_l\times(R+1)}$, $\widetilde{\boldsymbol{D}} = \left[\boldsymbol{D}, \boldsymbol{b}^\top\right] \in \mathbb{R}^{(R+1)\times M}$ be augmented input data and parameter matrices, respectively, to obtain (19)

$$\widetilde{\boldsymbol{D}}^* = \arg\min_{\widetilde{\boldsymbol{D}}} \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{X}^l}\widetilde{\boldsymbol{D}}\right\|_F^2 + \frac{\lambda}{\xi}\left\|\widetilde{\boldsymbol{D}}\right\|_{2,p}^p \tag{19}$$

As a result, the most effort for solving $\boldsymbol{D}$ and $\boldsymbol{b}$-subproblem is to handling $L_{2,p}$ norm of $\widetilde{\boldsymbol{D}}$.

Through some algebra, we have

$$\begin{aligned}\left\|\widetilde{\boldsymbol{D}}\right\|_{2,p}^p &= \sum_{i=1}^{R+1}\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}\right\|_2^p = \sum_{i=1}^{R+1}\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}\right\|_2^{p-2}\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}\right\|_2^2 \\ &= \sum_{i=1}^{R+1}\widetilde{\boldsymbol{D}}_{i,\cdot}\boldsymbol{Q}_{ii}\widetilde{\boldsymbol{D}}_{i,\cdot}^\top = Tr\left(\widetilde{\boldsymbol{D}}^\top\boldsymbol{Q}\widetilde{\boldsymbol{D}}\right)\end{aligned} \tag{20}$$

where $\boldsymbol{Q}$ is a $(R+1)\times(R+1)$ diagonal matrix whose diagonal entries are given as

$$\boldsymbol{Q}_{ii} = \left\|\widetilde{\boldsymbol{D}}_{i,\cdot}\right\|_2^{p-2}, \quad i = 1, 2, \cdots, R+1 \tag{21}$$

Although (19) is difficult to address directly, it can be solved efficiently in an alternative manner. To be specific, in the current iteration, the value of $\boldsymbol{Q}$ in the last iteration is known, We thereby can simplify (19) into (22)

$$\widetilde{\boldsymbol{D}}^* = \arg\min_{\widetilde{\boldsymbol{D}}} \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{X}^l}\widetilde{\boldsymbol{D}}\right\|_F^2 + \frac{\lambda}{\xi}\frac{p}{2}Tr\left(\widetilde{\boldsymbol{D}}^\top\boldsymbol{Q}\widetilde{\boldsymbol{D}}\right) \tag{22}$$

Furthermore, the feasibility of such transformation is guaranteed by Proposition 2.

[Proposition 2]: Define $f\left(\widetilde{\boldsymbol{D}}\right) = \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{X}^l}\widetilde{\boldsymbol{D}}\right\|_F^2 + \frac{\xi}{\lambda}\left\|\widetilde{\boldsymbol{D}}\right\|_{2,p}^p$ and let t indicate the iteration, if $\widetilde{\boldsymbol{D}}$ is optimized via (22), then f decreases monotonically, i.e., $f\left(\widetilde{\boldsymbol{D}}^{(t+1)}\right) \le f\left(\widetilde{\boldsymbol{D}}^{(t)}\right)$.

[Proof]: Firstly, let us introduce an useful inequality [43], i.e.,

$$\begin{aligned}&\frac{\left\|\boldsymbol{e}^{(t+1)}\right\|_2^p}{\left\|\boldsymbol{e}^{(t)}\right\|_2^p} - \frac{p}{2}\frac{\left\|\boldsymbol{e}^{(t+1)}\right\|_2^2}{\left\|\boldsymbol{e}^{(t)}\right\|_2^2} - 1 + \frac{p}{2} \le 0 \\ &\forall \boldsymbol{e}^{(t+1)}, \boldsymbol{e}^{(t)} \in \mathbb{R}^R, 0 < p < 2\end{aligned} \tag{23}$$

Denote $\boldsymbol{e}^{(t+1)} = \widetilde{\boldsymbol{D}}_{i,\cdot}^{(t+1)}$ and $\boldsymbol{e}^{(t)} = \widetilde{\boldsymbol{D}}_{i,\cdot}^{(t)}$ leads to

$$\begin{aligned}&\frac{p}{2}\sum_{i=1}^{R+1}\frac{\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}^{(t+1)}\right\|_2^2}{\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}^{(t)}\right\|_2^2}\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}^{(t)}\right\|_2^p \\ &\quad\ge \sum_{i=1}^{R+1}\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}^{(t+1)}\right\|_2^p - \sum_{i=1}^{R+1}\left(1-\frac{p}{2}\right)\left\|\widetilde{\boldsymbol{D}}_{i,\cdot}^{(t)}\right\|_2^p\end{aligned} \tag{24}$$

Subsequently, we can obtain

$$\begin{aligned}&\frac{p}{2}Tr\left(\widetilde{\boldsymbol{D}}^{(t+1)\top}\boldsymbol{Q}^{(t)}\widetilde{\boldsymbol{D}}^{(t+1)}\right) - \frac{p}{2}Tr\left(\widetilde{\boldsymbol{D}}^{(t)\top}\boldsymbol{Q}^{(t)}\widetilde{\boldsymbol{D}}^{(t)}\right) \\ &\quad\ge \left\|\widetilde{\boldsymbol{D}}^{(t+1)}\right\|_{2,p}^p - \left\|\widetilde{\boldsymbol{D}}^{(t)}\right\|_{2,p}^p\end{aligned} \tag{25}$$

According to (22), we have

$$\begin{aligned}&\left\|\boldsymbol{Y} - \widetilde{\boldsymbol{X}^l}\widetilde{\boldsymbol{D}}^{(t+1)}\right\|_F^2 + \frac{\lambda}{\xi}\frac{p}{2}Tr\left(\widetilde{\boldsymbol{D}}^{(t+1)\top}\boldsymbol{Q}^{(t)}\widetilde{\boldsymbol{D}}^{(t+1)}\right) \\ &\quad\le \left\|\boldsymbol{Y} - \widetilde{\boldsymbol{X}^l}\widetilde{\boldsymbol{D}}^{(t)}\right\|_F^2 + \frac{\lambda}{\xi}\frac{p}{2}Tr\left(\widetilde{\boldsymbol{D}}^{(t)\top}\boldsymbol{Q}^{(t)}\widetilde{\boldsymbol{D}}^{(t)}\right)\end{aligned} \tag{26}$$

Combing (25) and (26), we can complete this proof. □

Take the derivative of (22) with respective to $\widetilde{\boldsymbol{D}}$ and letting it to zero, we have

$$\begin{aligned}&\left(\widetilde{\boldsymbol{X}^l}\right)^\top\left(\boldsymbol{Y} - \widetilde{\boldsymbol{X}^l}\widetilde{\boldsymbol{D}}\right) + \frac{\lambda}{\xi}\frac{p}{2}\boldsymbol{Q}\widetilde{\boldsymbol{D}} = 0 \Rightarrow \\ &\widetilde{\boldsymbol{D}}^* = \left[\left(\widetilde{\boldsymbol{X}^l}\right)^\top\widetilde{\boldsymbol{X}^l} + \frac{\lambda}{\xi}\frac{p}{2}\boldsymbol{Q}\right]^{-1}\left(\widetilde{\boldsymbol{X}^l}\right)^\top\boldsymbol{Y}\end{aligned} \tag{27}$$

After obtaining the optimal $\widetilde{\boldsymbol{D}}^*$, $\boldsymbol{D}^*$ and $\boldsymbol{b}^*$'s optima can be determined accordingly.

4) $\boldsymbol{E}$-Subproblem: While fixing the other variables, it leads to

$$\begin{aligned}\boldsymbol{E}^* &= \arg\min_{\boldsymbol{E}} \frac{\lambda}{2}\|\boldsymbol{E}\|_* - Tr\left(\boldsymbol{W}^\top(\boldsymbol{B}-\boldsymbol{E})\right) + \frac{\mu}{2}\|\boldsymbol{B}-\boldsymbol{E}\|_F^2 \\ &= \arg\min_{\boldsymbol{E}} \frac{\lambda}{2}\|\boldsymbol{E}\|_* + \frac{\mu}{2}\left\|\boldsymbol{B}-\boldsymbol{E}-\frac{\boldsymbol{W}}{\mu}\right\|_F^2 - \frac{\mu}{2}\left\|\frac{\boldsymbol{W}}{\mu}\right\|_F^2 \\ &= \arg\min_{\boldsymbol{E}} \frac{\lambda}{2\mu}\|\boldsymbol{E}\|_* + \frac{1}{2}\left\|\boldsymbol{B}-\boldsymbol{E}-\frac{\boldsymbol{W}}{\mu}\right\|_F^2\end{aligned} \tag{28}$$

Proposition 3 shows how to solve $\boldsymbol{E}$-subproblem.

[Proposition 3]: Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{D\times R}$, and $\tau > 0$. Consider the following unconstrained optimization problem

$$\arg\min_{\boldsymbol{X}} = \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 + \tau\|\boldsymbol{X}\|_* \tag{29}$$

Suppose the singular value decomposition (SVD) of $\boldsymbol{Y}$ results in $\boldsymbol{Y} = \boldsymbol{U}diag(\boldsymbol{\sigma})\boldsymbol{V}^\top$, then $\boldsymbol{X}^* = \boldsymbol{U}diag\{\max(\boldsymbol{\sigma}-\tau, 0)\}\boldsymbol{V}^\top$. The proof details can be found in [54].

According to Proposition 3, $\boldsymbol{E}$'s optima can be determined as

$$\boldsymbol{E}^* = \boldsymbol{U}_2 diag\left\{\max\left(\boldsymbol{\sigma}_2 - \frac{\lambda}{2\mu}, 0\right)\right\}\boldsymbol{V}_2^\top \tag{30}$$

where $\boldsymbol{U}_2$, $\boldsymbol{V}_2$, and $\boldsymbol{\sigma}_2$ are calculated via performing SVD on $\boldsymbol{B} - \frac{\boldsymbol{W}}{\mu}$, i.e., $\boldsymbol{B} - \frac{\boldsymbol{W}}{\mu} = \boldsymbol{U}_2 diag(\boldsymbol{\sigma}_2)\boldsymbol{V}_2^\top$

For multiplier, it can be updated by

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \mu(\boldsymbol{B}-\boldsymbol{E}) \tag{31}$$

When the optimal parameters are determined, the PCSELM is available to predict output value, when a query sample $\boldsymbol{z}_*$ comes. Here, the derived soft sensor's estimation $\hat{\boldsymbol{y}}_*$ of $\boldsymbol{y}_*$ is obtained by

$$\hat{\boldsymbol{y}}_* = \boldsymbol{D}^\top\boldsymbol{B}^\top\boldsymbol{x}_* + \boldsymbol{b} \tag{32}$$

C. Algorithm Analysis

The procedure of the PCSELM is shown in Algorithm 1 and its theoretical analysis and convergence is guaranteed by the following rigorous proof.

[Proposition 4]: The objective function $J(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{b})$ in (10) is bounded from below and monotonically decreases with each optimization step for each parameters, and therefore it converges.

[Proof]: Since $J(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{b})$ is the summation of norms, we have $J(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{b}) \ge 0$ for any $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$, and $\boldsymbol{b}$. Then it is bounded from below. Based on the Propositions 1-3

Algorithm 1 PCSELM

Input: Labeled dataset $\{\boldsymbol{Z}^l, \boldsymbol{Y}\}$, unlabeled dataset $\boldsymbol{Z}^u$, hyper-parameters λ, ξ and p
implement random nonlinear mapping
Initialize $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{b}$, $\boldsymbol{E}$ and $\boldsymbol{W}$
Set $\mu = 0.1$
while stopping criterion is not met **do**
Compute $\boldsymbol{A}$ by (15)
Compute $\boldsymbol{B}$ by (18)
Compute $\boldsymbol{D}$ and $\boldsymbol{b}$ by (27)
Compute $\boldsymbol{E}$ by (30)
Update $\boldsymbol{W}$ by (31)
Update μ by $\mu \leftarrow \mu \times 1.1$
end while
Output: Optimal $\boldsymbol{B}$, $\boldsymbol{D}$, and $\boldsymbol{b}$

and the gradient descend property, one can conclude that the objective function decreases monotonically, which indicates that it converges according to the monotone convergence theorem. □

Next, we analyze the computational complexity of PCSELM. Algorithm 1 is comprised of two loops. The former is to update all parameters in PCSELM, while the latter is for $\boldsymbol{B}$'s gradient descend. Assume the maximum iteration count for the two loops are K and T, respectively. The computational costs of calculating $\boldsymbol{A}$, $\boldsymbol{D}$, and $\boldsymbol{E}$ once are $O(R^3)$, $O(R^3)$, and $O(R^5)$. While $\boldsymbol{B}$'s computational cost is $O(KN^3)$. Thus, the total computational complexity of the proposed method is $O(TKN^3 + TR^5 + 2TR^3)$.

IV. Case Study

In this section, the performance of the PCSELM is evaluated using an industrial process. The performance of the proposed model is compared with other soft sensor models. To quantitatively evaluate prediction performance of these soft-sensors, the root mean square error ($RMSE$) and determination (R^2) are employed. These two static indexes are formulated as:

$$RMSE = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \hat{y}_i)^2} \tag{33}$$

$$R^2 = 1 - \sum_{i=1}^{N_t}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N_t}(y_i - \bar{y})^2 \tag{34}$$

where N_t means the testing sample count, y_i is the real value of output variable y at the i-th sampling point, $\hat{y}_i$ is its prediction, $\bar{y}$ is the mean of output values. Both these indexes reveal the dispersion degree of the predictive result. A smaller $RMSE$ and a larger R^2 suggest a better soft-sensing performance. The configuration of the computer used for conducting experiments is as follows. CPU: AMD Ryzen 7 5800H (3.20 GHz), OS: Windows 11 (64 bit), RAM: 16GB.

A. Description of the Debutanizer Column Process

In the refine process of naphtha split, a debutanizer column is pivotal equipment, which is mainly used to separate butane from naphtha stream [55]. For the purpose of improving the control quality of a debutanizer column, it is of great importance to detect the butane content in time. However, the measurement of butane content is not as easy as that of some process variables like temperature and pressure, because it takes incorporates plenty of time and human efforts for content analysis. So it is an alternative to estimate the butane content via virtual sensing in such a situation. Seven variables which have strong connections to the butane content have been chosen for the virtual sensor development. The flowchart of the debutanizer column is shown in Fig. 2, and the variables are listed in Table I.

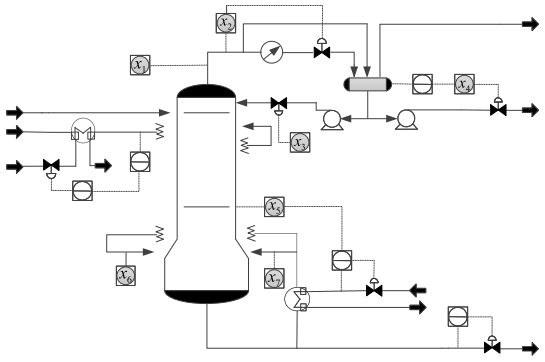


Fig. 2. Flowchart of the debutanizer column [55].

TABLE I
The Debutanizer Distillation Process Variables

Tags	Descriptions
u_1	Top temperature
u_2	Top pressure
u_3	Reflux flow
u_4	Flow to next process
u_5	Sixth tray temperature
u_6	Bottom temperature A
u_7	Bottom temperature B
y	Butane C4 content in IC5

B. Results and Discussions

For accurate estimation of the outputs, the process dynamics are taken into consideration [56]. That means, new variables are augmented by a number of previous samples with lagged input and output data. To simulate the situation of semi-supervision, different labeling rates (LR), i.e., the proportion of labeled samples in a training dataset, ranging from 10% to 50% are set for prediction performance evaluation. Since the smallest LR is 10%, the lagged time of the output data is set as 10. To this end, the augmented variables $\boldsymbol{x}(k)$ that are used for soft sensor modeling are designed as

$$\boldsymbol{x}(k) = \begin{bmatrix} u_1(k), u_2(k), u_3(k), u_4(k), u_5(k), u_5(k-1), \\ u_5(k-2), u_5(k-3), (u_6(k)+u_7(k))/2, y(k-10) \end{bmatrix}^T \tag{35}$$

Hence, 10 variables are used for predicting the output y at sampling time k. The debutanizer distillation dataset is a benchmark dataset to validate different approaches in the field of process monitoring. A number of 2384 input and output data samples are used in this paper. For model construction and evaluation, $\frac{1}{2}$ samples are used for model

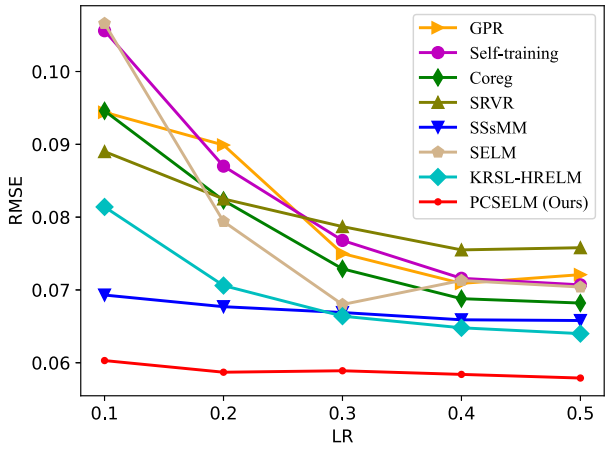


Fig. 3. Testing $RMSE$ of each model.

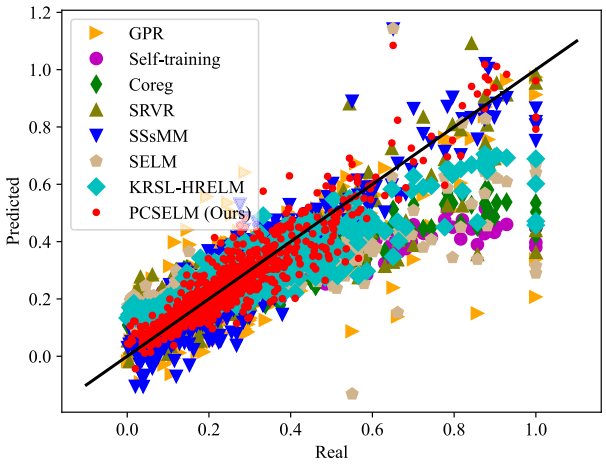


Fig. 4. Scatter plot comparisons among each method.

training, $\frac{1}{6}$ are used for validating dataset, and the remaining $\frac{1}{3}$ are used for testing. In order to exhibit the advantages of the PCSELM, several relevant soft sensing approaches are also performed for comparison. The benchmark ones include Gaussian process regression (GPR), self-training KNN, coreg [46], semi-supervised relevant vector machine (SRVR) [51], semi-supervised ELM (SELM) [31], semi-supervised Student's t mixture model (SsSMM) [52], and semi-supervised Kernel Risk-Sensitive Loss Based Hyper-graph Regularized Robust Extreme Learning Machine (KRSL-HRELM) [53]. The model settings of PCSELM are provided as follows. The tanh function is chosen as the activation function, the number of hidden nodes is 50, the number of remaining principal components are 25, and the model coefficients are set as $\xi = 10^{-4}$, $\lambda = 10^{-3}$, $p = 1.9$.

The $RMSE$ and R^2 values among the seven soft sensors with different ratios of labeled training samples are tabulated in Table II. The values of quantitative indices listed in Table II confirm the predictive advantages of the PCSELM-based soft sensor over those based on its peers. Since in each scenario with a certain LR, PCSELM gives the lowest $RMSE$ and the highest R^2 among the seven methods. We can conclude that when both labeled and unlabeled samples are used for training, the performance of PCSELM illustrates that it can provide superior predictions than the benchmark ones.

More intuitively, $RMSE$ values are displayed in Fig. 3. From it, we can find that the $RMSE$ value becomes smaller and the R^2 value becomes larger with the increase of LR, which indicates that the prediction performance can be improved by increasing the labeled data samples. Conversely, the predictive accuracies basically decline as LR decreases. However, the performance deterioration for the PCSELM is less than those for the other ones. Specifically, as LR decreases from 50% to 10%, the performance deteriorations are 30.9%, 49.4%, 38.7%, 17.4%, 5.3%, 51.4%, 27.2%, 4.1% for each method. This demonstrates that the proposed PCSELM model is insensitive to decreasing the amount of labeled data, thus it is applicable to the scenario where the number of labeled samples is quite small.

For visual comparison of the prediction results, their scatter plots with LR=10% are displayed in Fig. 4. In Fig.4, the

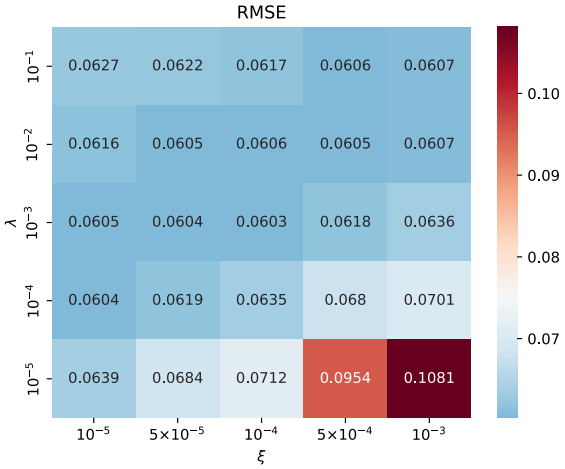


Fig. 5. $RMSE$ variations with different values of ξ and λ.

horizontal axis and the vertical axis are the real and predicted values, respectively. That means the closer the scatter is to the diagonal line, the better the prediction accuracy is. PCSELM illustrates superior performances than other semi-supervised approaches, since Fig. 4 manifests that the closeness degree of the scatters around the black line by the PCSELM are noticeably better than those by the other soft sensing approaches with the same amount of labeled samples utilized for model training.

There exist two trade-off hyper-parameters ξ and λ in PCSELM. To illustrate influence of these two hyper-parameters on prediction performance, comparison experiments are carried out by tuning ξ and λ. Figs. 5 and 6 show the detailed $RMSE$ and R^2 values with various ξ and λ when LR=10%. From these comparison results, one can observe that the larger ξ or smaller λ would lead to prediction performance deterioration. It can also be found that the smaller $RMSE$ and larger R^2 values will be gained when ξ and λ are in the middle of their ranges. Consequently, we select $\xi = 10^{-4}$ and $\lambda = 10^{-3}$ as the final hyper-parameters for soft sensor modeling.

TABLE II
PREDICTION RESULTS FOR THE DEBUTANIZER COLUMN BY DIFFERENT METHODS

Method	Dataset	LR=10%		LR=20%		LR=30%		LR=40%		LR=50%	
		$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2
GPR	validating	0.0963	0.532	0.0929	0.578	0.0788	0.695	0.0711	0.762	0.0709	0.759
	testing	0.0944	0.578	0.0899	0.626	0.0750	0.745	0.0709	0.774	0.0721	0.758
Self-training	validating	0.1065	-1.303	0.0876	0.153	0.0774	0.475	0.0739	0.556	0.0727	0.579
	testing	0.1056	-1.178	0.0870	0.197	0.0768	0.510	0.0716	0.605	0.0707	0.621
Coreg[46]	validating	0.0956	-0.335	0.0819	0.354	0.0728	0.578	0.0670	0.628	0.0677	0.658
	testing	0.0946	-0.240	0.0823	0.357	0.0729	0.593	0.0688	0.661	0.0682	0.662
SRVR[51]	validating	0.0879	0.462	0.0799	0.577	0.0789	0.555	0.0772	0.589	0.0767	0.595
	testing	0.0890	0.446	0.0825	0.566	0.0787	0.564	0.0755	0.602	0.0758	0.597
SsSMM[52]	validating	0.0656	0.824	0.0625	0.824	0.0617	0.825	0.0610	0.829	0.0613	0.827
	testing	0.0693	0.820	0.0677	0.814	0.0669	0.815	0.0659	0.821	0.0658	0.821
SELM[31]	validating	0.1054	-0.693	0.0793	0.406	0.0668	0.670	0.0624	0.731	0.0609	0.756
	testing	0.1066	-0.550	0.0794	0.491	0.0680	0.691	0.0713	0.715	0.0704	0.735
KRSL-RELM[53]	validating	0.0831	0.287	0.0728	0.572	0.0684	0.670	0.0671	0.699	0.0665	0.714
	testing	0.0814	0.330	0.0706	0.603	0.0664	0.695	0.0648	0.725	0.0640	0.742
PCSELM(Ours)	validating	0.0631	0.838	0.0613	0.838	0.0609	0.835	0.0602	0.837	0.0602	0.835
	testing	**0.0603**	**0.857**	**0.0587**	**0.857**	**0.0589**	**0.852**	**0.0584**	**0.853**	**0.0579**	**0.854**

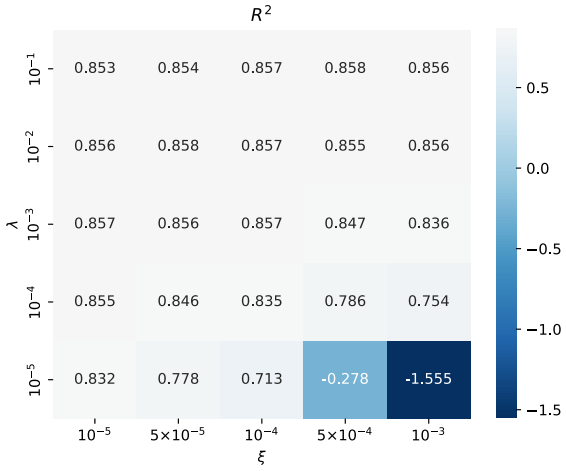


Fig. 6. R^2 variations with different values of ξ and λ.

Besides, we provide an ablation study to verify the contribution of the parameter regularizations (i.e., $J_{P_1} + J_{P_2}$). The ablation experiment is divided into two groups including PCSELM-0 and PCSELM-1, where the former represents the method with the parameter regularizations and the latter represents the method without such regularizations. The comparison results are shown in Table III. We can readily see that the prediction performances of PCSELM-0 are more satisfactory than those of PCSELM-1, which illustrates the significance of the parameter regularizations. This is also reflected in Figs. 5 and 6. From Figs. 5 and 6, it can be seen that decreasing the value of λ will degrade the accuracy of the PCSELM model, which also indicates that the parameter regularizations contribute to satisfactory prediction performance.

Detailed predictions of the Butane content on the testing samples by three ELM-based methods with LR=10% are presented in Figs. 7, 8, and 9, where the real values are drawn in blue, and the predicted ones are drawn in red. And

TABLE III
$RMSE$ AND R^2 VALUES OF THE ABLATION EXPERIMENT

LR	PCSELM-0		PCSELM-1	
	$RMSE$	R^2	$RMSE$	R^2
10%	0.0603	0.857	0.1210	-4.261
20%	0.0587	0.857	0.1020	-0.866
30%	0.0589	0.852	0.0904	0.0126
40%	0.0584	0.853	0.0868	0.195
50%	0.0579	0.854	0.0844	0.284

their errors are displayed in Fig. 10. As is observed from them, predictive plots achieved by the KRSL-HRELM and PCSELM can basically capture the time trend variations of the Butane content, and their predictions significantly fluctuate. In contrast, the SELM's predictions are much more smooth. Besides, the predicted line of the PCSELM tracks the real line pretty well, especially for the large Butane content values (around the 350th and 650th testing samples). This is because PCSELM can take good care of both the dimension reduction and semi-supervised regression modeling issue, such that the essential information of the data structure can be better extracted. In contrast with the three ELM-based soft sensors, the improvement yielded by the PCSELM can be regarded as significant.

Statistical test has been carried out for testing if the performance improvement by the PCSELM is significant. The nonparametric Wilcoxon test is adopted to test if the squared predicted errors obtained by two methods are significantly different. Six null-hypotheses given the significance level $\alpha = 0.05$, the decisions of the Wilcoxon test are summarized in Table IV. As can be seen from it, the differences between the performance of the PCSELM and the other methods are significant with the low proportion of the labeled data samples.

Some specific analysis are done to assess the model residuals. Fig. 11 provides the four-plot results for PCSELM, which include a sequence plot, a histogram, a normal probability test,

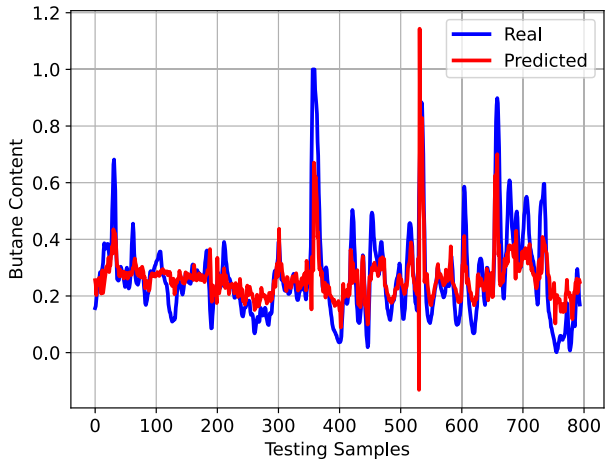


Fig. 7. Predicted plots of Butane content by the SELM model.

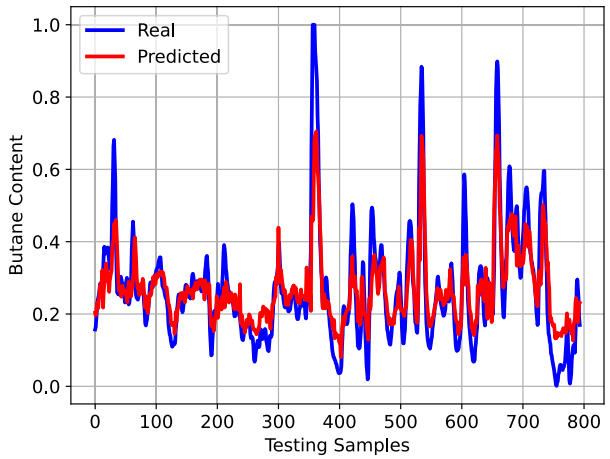


Fig. 8. Predicted plots of Butane content by the KRSL-RELM model.

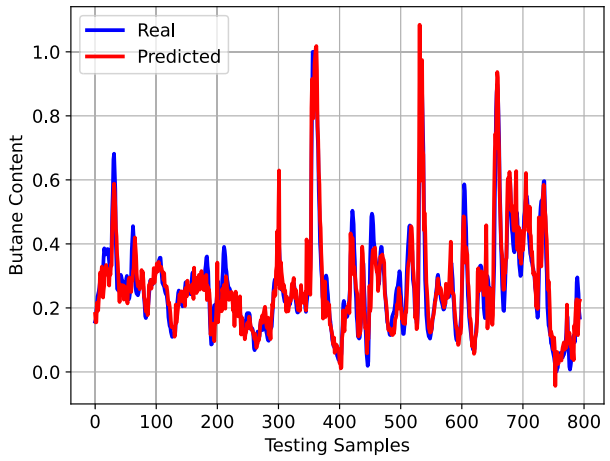


Fig. 9. Predicted plots of Butane content by the PCSELM model.

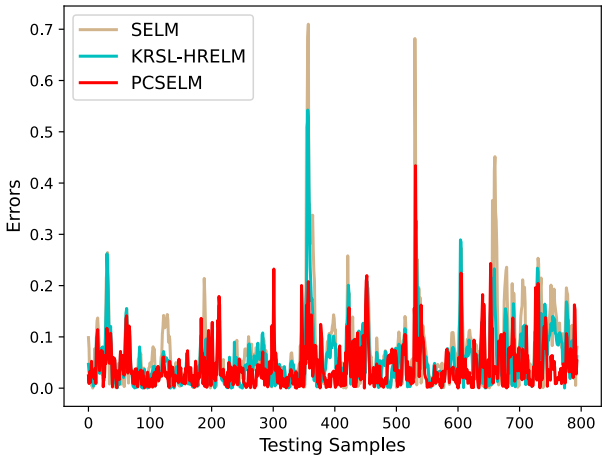
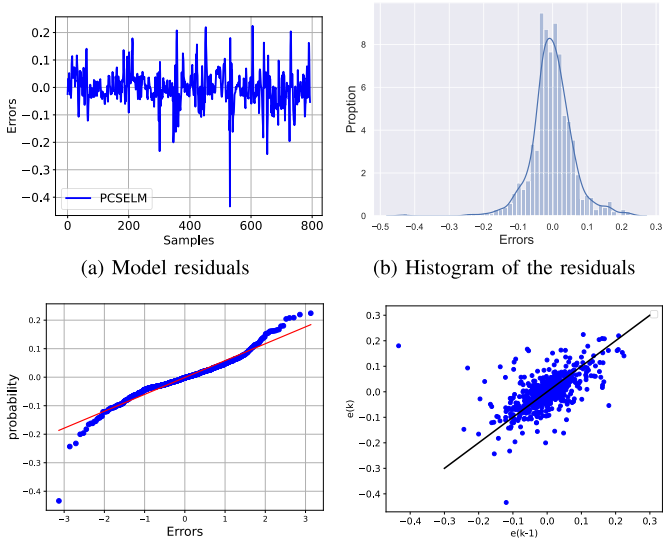


Fig. 10. Errors of three ELM-based methods.

TABLE IV
STATISTICAL TEST FOR THE SQUARED ERRORS

$\mathcal{H}_0$	LR				
	10%	20%	30%	40%	50%
$E^2_{Ours} = E^2_{GPR}$	reject	reject	accept	reject	reject
$E^2_{Ours} = E^2_{Self-training}$	reject	reject	reject	accept	accept
$E^2_{Ours} = E^2_{Coreg}$	reject	reject	accept	accept	accept
$E^2_{Ours} = E^2_{SRVR}$	reject	reject	reject	reject	reject
$E^2_{Ours} = E^2_{SsSMM}$	reject	reject	reject	reject	reject
$E^2_{Ours} = E^2_{SELM}$	reject	reject	reject	reject	reject
$E^2_{Ours} = E^2_{KRSL-RELM}$	reject	reject	reject	reject	reject



(a) Model residuals (b) Histogram of the residuals

(c) Quantile-quantile plot (d) Lag

Fig. 11. Four plots of the prediction errors on the testing dataset.

and a lag plot. From the detailed error sequence plot, most of the prediction errors are maintained around zero. Only a few errors' absolute values are larger than 0.1. Therefore, the prediction of the proposed PCSELM is accurate for the major testing data. From the normal probability plot, we can see that the most prediction errors roughly follow a normal distribution because their closeness to the red reference line. This can also be validated by the histogram plot since the fitted line is nearly bell-shaped. From the scatter lag plot between each sample's prediction error and that of its previous lagged sample. There exits a approximately linear pattern between them. It indicates that the debutanizer column is a dynamic process, that is why we take lagged process variables for soft sensing. According to the comparisons and analyzes of the industrial case study, we can conclude that the proposed PCSELM is an efficient and flexible tool for semi-supervised soft sensing.

V. Conclusion

In the current work, for the purpose of quality prediction in the scenario of insufficient labeled samples, we present a principal component based semi-supervised extreme learning machine (PCSELM) as well as its corresponding training algorithm. Theoretical and experimental analysis demonstrate the advantages of the PCSELM, such as (1) improving the generalization performance by exploiting the low-dimensional principal components; (2) the capability for the semi-supervised learning; and (3) the theoretically guaranteed convergence for the parameter optimization algorithm. Comprehensive experiment results are reported for an industrial process, which indicate the effectiveness of the proposed PCSELM.

Potential future works can be done to integrate the multimodality information by using the cluster-based multiple models and just in time learning strategy. Besides, exploring the other optimization strategies like accelerated proximal gradient descent [57] approach would be an interesting topic, which deserves deep investigation in the future.

Acknowledgment

The authors are grateful for the efforts of their colleagues at the Sino-German Center of Intelligent Systems, Tongji University.

References

[1] S. Yao, Q. Kang, M. Zhou, M. J. Rawa, and A. Albeshri, "Discriminative manifold distribution alignment for domain adaptation," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 53, no. 2, pp. 1183–1197, Feb. 2023.

[2] S. Yao, Q. Kang, M. Zhou, M. J. Rawa, and A. Abusorrah, "A survey of transfer learning for machinery diagnostics and prognostics," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 2871–2922, Apr. 2023, doi: 10.1007/s10462-022-10230-4.

[3] Q. Deng, Q. Kang, L. Zhang, M. Zhou, and J. An, "Objective space-based population generation to accelerate evolutionary algorithms for large-scale many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 27, no. 2, pp. 326–340, Apr. 2023, doi: 10.1109/TEVC.2022.3166815.

[4] X. Wang, Q. Kang, M. Zhou, S. Yao, and A. Abusorrah, "Domain adaptation multitask optimization," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4567–4578, Jul. 2023, doi: 10.1109/TCYB.2022.3222101.

[5] L. Feng and C. Zhao, "Fault description based attribute transfer for zero-sample industrial fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1852–1862, Mar. 2021.

[6] L. Zeng and Z. Ge, "Bayesian network for dynamic variable structure learning and transfer modeling of probabilistic soft sensor," *J. Process Control*, vol. 100, pp. 20–29, Apr. 2021.

[7] L. Feng, C. Zhao, and Y. Sun, "Dual attention-based encoder–decoder: A customized sequence-to-sequence learning for soft sensor development," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3306–3317, Aug. 2021.

[8] X. Kong and Z. Ge, "Deep PLS: A lightweight deep learning model for interpretable and efficient data analytics," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 11, 2022, doi: 10.1109/TNNLS.2022.3154090.

[9] X. Yuan, Y. Gu, Y. Wang, C. Yang, and W. Gui, "A deep supervised learning framework for data-driven soft sensor modeling of industrial processes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4737–4746, Nov. 2020.

[10] X. Shi, Q. Kang, M. Zhou, A. Abusorrah, and J. An, "Soft sensing of nonlinear and multimode processes based on semi-supervised weighted Gaussian regression," *IEEE Sensors J.*, vol. 20, no. 21, pp. 12950–12960, Nov. 2020.

[11] R. Jiao, K. Peng, and J. Dong, "Remaining useful life prediction for a roller in a hot strip mill based on deep recurrent neural networks," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 7, pp. 1345–1354, Jul. 2021.

[12] C. Ou et al., "Quality-driven regularization for deep learning networks and its application to industrial soft sensors," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 18, 2022, doi: 10.1109/TNNLS.2022.3144162.

[13] 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 Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7682–7694, Dec. 2022, doi: 10.1109/TNNLS.2021.3086323.

[14] Q. Sun and Z. Ge, "A survey on deep learning for data-driven soft sensors," *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 5853–5866, Sep. 2021.

[15] X. Shi, Q. Kang, J. An, and M. Zhou, "Novel L1 regularized extreme learning machine for soft-sensing of an industrial process," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1009–1017, Feb. 2022.

[16] J. Bi, X. Zhang, H. Yuan, J. Zhang, and M. Zhou, "A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1869–1879, Jul. 2022.

[17] S. Liao, X. Jiang, and Z. Ge, "Weakly supervised multilayer perceptron for industrial fault classification with inaccurate and incomplete labels," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 1192–1201, Apr. 2022.

[18] Z. Yao and C. Zhao, "FIGAN: A missing industrial data imputation method customized for soft sensor application," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 3712–3722, Oct. 2022, doi: 10.1109/TASE.2021.3132037.

[19] C. Zhao, J. Chen, and H. Jing, "Condition-driven data analytics and monitoring for wide-range nonstationary and transient continuous processes," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 4, pp. 1563–1574, Oct. 2021.

[20] Z. Chai and C. Zhao, "A fine-grained adversarial network method for cross-domain industrial fault diagnosis," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 3, pp. 1432–1442, Jul. 2020.

[21] Y. Liu and M. Xie, "Rebooting data-driven soft-sensors in process industries: A review of kernel methods," *J. Process Control*, vol. 89, pp. 58–73, May 2020.

[22] L. Wang, J. Zeng, X. Liang, Y. He, S. Luo, and J. Cai, "Soft sensing of a nonlinear multimode process using a self organizing model and conditional probability density analysis," *Ind. Eng. Chem. Res.*, vol. 58, no. 31, pp. 14267–14274, Aug. 2019.

[23] J. Zhu, Z. Ge, and Z. Song, "Robust supervised probabilistic principal component analysis model for soft sensing of key process variables," *Chem. Eng. Sci.*, vol. 122, pp. 573–584, Jan. 2015.

[24] H. Kodamana, R. Raveendran, and B. Huang, "Mixtures of probabilistic PCA with common structure latent bases for process monitoring," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 2, pp. 838–846, Mar. 2019.

[25] H. Liu, C. Yang, B. Carlsson, S. J. Qin, and C. Yoo, "Dynamic nonlinear partial least squares modeling using Gaussian process regression," *Ind. Eng. Chem. Res.*, vol. 58, no. 36, pp. 16676–16686, Sep. 2019.

[26] X. Li, F. Wu, R. Zhang, and F. Gao, "Nonlinear multivariate quality prediction based on OSC-SVM-PLS," *Ind. Eng. Chem. Res.*, vol. 58, NO. 19, pp. 8154–8161, 2019.

[27] Q. Zhu, Q. Liu, and S. J. Qin, "Concurrent monitoring and diagnosis of process and quality faults with canonical correlation analysis," *J. Process Control*, vol. 60, pp. 95–103, Jul. 2017.

[28] X. Yang, W. Liu, W. Liu, and D. Tao, "A survey on canonical correlation analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2349–2368, Jun. 2021.

[29] W. Shao, Z. Ge, Z. Song, and K. Wang, "Nonlinear industrial soft sensor development based on semi-supervised probabilistic mixture of extreme learning machines," *Control. Eng. Pract.*, vol. 91, Oct. 2019, Art. no. 104098.

[30] Z. Geng, J. Dong, J. Chen, and Y. Han, "A new self-organizing extreme learning machine soft sensor model and its applications in complicated chemical processes," *Eng. Appl. Artif. Intell.*, vol. 62, pp. 38–50, Jun. 2017.

[31] G. Huang, S. Song, J. N. D. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning machines," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2405–2417, Dec. 2014.

[32] G. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[33] X. Zhang, X. Deng, and P. Wang, "Double-level locally weighted extreme learning machine for soft sensor modeling of complex nonlinear industrial processes," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1897–1905, Jan. 2021.

[34] T. Ouyang et al., "NOx measurements in vehicle exhaust using advanced deep ELM networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.
[35] L. Guo, R. Li, and B. Jiang, "An ensemble broad learning scheme for semisupervised vehicle type classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5287–5297, Dec. 2021.
[36] J. Lu, J. Ding, C. Liu, and T. Chai, "Hierarchical-Bayesian-based sparse stochastic configuration networks for construction of prediction intervals," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3560–3571, Aug. 2022, doi: 10.1109/TNNLS.2021.3053306.
[37] C. Wei, W. Shao, and Z. Song, "Virtual sensor development for multioutput nonlinear processes based on bilinear neighborhood preserving regression model with localized construction," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2500–2510, Apr. 2021.
[38] K. Wang, X. Yuan, J. Chen, and Y. Wang, "Supervised and semi-supervised probabilistic learning with deep neural networks for concurrent process-quality monitoring," *Neural Netw.*, vol. 136, pp. 54–62, Apr. 2021.
[39] X. Xiu, Y. Yang, L. Kong, and W. Liu, "Laplacian regularized robust principal component analysis for process monitoring," *J. Process Control*, vol. 92, pp. 212–219, Aug. 2020.
[40] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, Jan. 2006.
[41] X. Zhen, M. Yu, X. He, and S. Li, "Multi-target regression via robust low-rank learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 497–504, Feb. 2018.
[42] Q. Wang, Q. Gao, G. Sun, and C. Ding, "Double robust principal component analysis," *Neurocomputing*, vol. 391, pp. 119–128, May 2020.
[43] Q. Wang, Q. Gao, X. Gao, and F. Nie, "$\ell_{2,p}$-norm based PCA for image recognition," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1336–1346, 2018.
[44] X. Shi and W. Xiong, "Adaptive ensemble learning strategy for semi-supervised soft sensing," *J. Franklin Inst.*, vol. 357, no. 6, pp. 3753–3770, Apr. 2020.
[45] X. Shi and W. Xiong, "Approximate linear dependence criteria with active learning for smart soft sensor design," *Chemometrics Intell. Lab. Syst.*, vol. 180, pp. 88–95, Sep. 2018.
[46] Z. Zhou and M. Li, "Semi-supervised regression with Co-training," in *Proc. IJCAI*, 2005, pp. 908–916.
[47] J. Wang, F. Xie, F. Nie, and X. Li, "Robust supervised and semisupervised least squares regression using $\ell_{2,p}$-norm minimization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 23, 2022, doi: 10.1109/TNNLS.2022.3150102.
[48] Z. Ge, Z. Song, and F. Gao, "Self-training statistical quality prediction of batch processes with limited quality data," *Ind. Eng. Chem. Res.*, vol. 52, no. 2, pp. 979–984, Jan. 2013.
[49] D. Li, Y. Liu, and D. Huang, "Development of semi-supervised multiple-output soft-sensors with Co-training and tri-training MPLS and MRVM," *Chemometrics Intell. Lab. Syst.*, vol. 199, Apr. 2020, Art. no. 103970.
[50] L. Feng, C. Zhao, and B. Huang, "Adversarial smoothing tri-regression for robust semi-supervised industrial soft sensor," *J. Process Control*, vol. 108, pp. 86–97, Dec. 2021.
[51] Y. Sun, X. Liu, and Z. Zhang, "Quality prediction via semisupervised Bayesian regression with application to propylene polymerization," *J. Chemometrics*, vol. 32, no. 10, p. e3052, Oct. 2018.
[52] W. Shao, Z. Ge, Z. Song, and J. Wang, "Semisupervised robust modeling of multimode industrial processes for quality variable prediction based on Student's t mixture model," *IEEE Trans. Ind. Informat.*, vol. 16, no. 5, pp. 2965–2976, May 2020.
[53] L.-R. Ren, J.-X. Liu, Y.-L. Gao, X.-Z. Kong, and C.-H. Zheng, "Kernel risk-sensitive loss based hyper-graph regularized robust extreme learning machine and its semi-supervised extension for classification," *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107226.
[54] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.
[55] L. Fortuna, S. Graziani, A. Rizzo, and M. G. Xibilia, *Soft Sensors For Monitoring and Control of Industrial Processes*. Cham, Switzerland: Springer, 2007.
[56] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3235–3243, Jul. 2018.
[57] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2013.

Xudong Shi (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and its automation and the M.S. degree in control theory and control engineering from Jiangnan University, Wuxi, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in control theory and control engineering with Tongji University, Shanghai, China. His research interests include machine learning, industrial process monitoring, and soft sensor development.

Qi Kang (Senior Member, IEEE) received the B.S. degree in automatic control and the M.S. and Ph.D. degrees in control theory and control engineering from Tongji University, Shanghai, China, in 2002, 2005, and 2009, respectively.

From 2007 to 2008, he was a Research Associate with the University of Illinois at Chicago, Chicago, IL, USA. From 2014 to 2015, he was a Visiting Scholar with the New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Professor with the Department of Control Science and Engineering and the Shanghai Institute of Intelligent Science and Technology, Tongji University. His research interests include swarm intelligence, evolutionary computation, machine learning, and intelligent control and optimization.

Dr. Kang was the General Chair of the 19th IEEE International Conference on Networking, Sensing and Control (ICNSC 2022). He is also an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS. He is the Secretary General with Shanghai Association for Systems Simulation.

Hanqiu Bao (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and its automation from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016, and the M.S. degree in control theory and control engineering from the Shanghai Institute of Technology, Shanghai, China, in 2019. He is currently pursuing the Ph.D. degree in control theory and control engineering with Tongji University, Shanghai. His research interests include MPC, underactuated system control, and nonlinear systems modeling.

Wangya Huang received the M.S. degree from the Pohang University of Science and Technology in 2005. He is currently pursuing the Ph.D. degree in electronic information with Tongji University, Shanghai, China. He is also a Chief Engineer with the Silicon Steel Business Unit, Baoshan Iron & Steel Company Ltd. His research interests include industrial big data and intelligent decision making systems.

Jing An (Member, IEEE) received the B.S. degree in automatic control, the M.S. degree in traffic information engineering and control, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 2002, 2005, and 2013, respectively. She is currently an Associate Professor with the School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai. Her research interests include computational intelligence, multi-objective optimization, and intelligent information processing.