SCIENCE CHINA
Information Sciences

• LETTER •

April 2020, Vol. 63 149205:1–149205:3
https://doi.org/10.1007/s11432-018-9624-8

Multi-rate principal component regression model for soft sensor application in industrial processes

Le ZHOU[1,2,3], Yaoxin WANG[2] & Zhiqiang GE[1*]

[1] *State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou* 310027*, China;*
[2] *School of Automation and Electrical Engineering, Zhejiang University of Science and Technology, Hangzhou* 310023*, China;*
[3] *Hangzhou SIASUN Robot and Automation Co., LTD., Hangzhou* 311225*, China*

Received 12 August 2018/Accepted 15 October 2018/Published online 25 September 2019

Citation Zhou L, Wang Y X, Ge Z Q. Multi-rate principal component regression model for soft sensor application in industrial processes. Sci China Inf Sci, 2020, 63(4): 149205, https://doi.org/10.1007/s11432-018-9624-8

Dear editor,
For obtaining the accurate and timely information from the modern process industries, an increasing number of online hardware sensors are equipped for process monitoring and control, energy management and environmental protection purpose [1, 2]. However, some key variables of the process cannot be measured by these online sensors, which requires the off-line laboratory analysis. For solving these problems, an inferential model called soft sensor is usually utilized [3].

Most traditional soft sensor methods are designed under the assumption that the sampling rate of the process and quality variables is the same [4]. In most chemical processes, the sampling rates of the process and quality variables may vary among a large range (from 1 s to 24 h). To integrate the multi-rate measurements for soft sensor modeling, the semi-supervised methods are developed in recent years. Zhou et al. [5] used a semi-supervised probabilistic latent variable regression model (SSPLVR) for process monitoring in both continuous and batch process. Yao et al. [6] proposed a weighted latent factor analysis model using the measurements in a dual-rate system. Shao et al. [7] proposed a novel semi-supervised selective ensemble learning soft sensor model where the distance to model method is defined as the criterion.

Even though the semi-supervised models are efficient in these cases, they are all used in a dual-rate process and it is not straightforward to extend them to the multi-rate processes. Hence, it is desirable to develop a soft senor model using the compelete multi-rate measurements without down-sampling or up-sampling. In this study, a multi-rate principal component regression model (MRPCR) is proposed for quality prediction purpose using the multi-rate samples.

Model and methodology. Consider an industrial process with N kinds of sampling rates for the process variables and S kinds of sampling rates for the quality indicators, which are noted as $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ and $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_S\}$, the multi-rate principal component regression model is given as

$$\begin{cases} \boldsymbol{x}_n{=}\boldsymbol{\Phi}_n\boldsymbol{t}{+}\boldsymbol{\varepsilon}_n, \ n=1,2,\ldots,N, \\ \boldsymbol{y}_s{=}\boldsymbol{\Psi}_s\boldsymbol{t}{+}\boldsymbol{\xi}_s, \ s=1,2,\ldots,S, \end{cases} \tag{1}$$

where $\boldsymbol{X}_1 \in \mathbb{R}^{K_1\times M_1}, \boldsymbol{X}_2 \in \mathbb{R}^{K_2\times M_2}, \ldots, \boldsymbol{X}_N \in \mathbb{R}^{K_N\times M_N}$. It indicates that sample numbers of the process variables with different sampling rates are also diverse, which is written as $\boldsymbol{X}_n = \{\boldsymbol{x}_{n1}, \boldsymbol{x}_{n2}, \ldots, \boldsymbol{x}_{nK_n}\}, n = 1,2,3,\ldots,N$. Similarly, the quality variables are given as $\boldsymbol{Y}_1 \in \mathbb{R}^{J_1\times H_1}, \boldsymbol{Y}_2 \in \mathbb{R}^{J_2\times H_2}, \ldots, \boldsymbol{Y}_S \in \mathbb{R}^{J_S\times H_S}$ and $\boldsymbol{Y}_s = \{\boldsymbol{y}_{s1}, \boldsymbol{y}_{s2}, \ldots, \boldsymbol{y}_{sJ_s}\}, s = 1,2,3,\ldots,S$. The key factor of the MRPCR model is the latent variable $\boldsymbol{T}{=}\ \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_K\} \in \mathbb{R}^{K\times D}$, where the total sampling time interval is K and it is guaranteed that at least one kind of sampling-rate variable is collected at each sampling inter-

* Corresponding author (email: gezhiqiang@zju.edu.cn)

val. Hence, it is readily to obtain that $K \geqslant \max\{K_1, K_2, \ldots, K_N, J_1, J_2, \ldots, J_S\}$. The latent variable $\boldsymbol{t}$ is determined and shared by all the multi-rate measurements. The loading matrixes are $\boldsymbol{\Phi}_n \in \mathbb{R}^{M_n \times D}, n = 1, 2, \ldots, N$ and $\boldsymbol{\Psi}_s \in \mathbb{R}^{M_s \times D}, s = 1, 2, \ldots, S$, respectively. The measurement noises are $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_s$.

The model inferences of the proposed MRPCR can be made in the probabilistic framework. Usually, we assumed that the prior distribution of the latent variable follows the Gaussian distribution with zero mean and unit variance and the measurement noises follow the isotropic Gaussian distributions as $\boldsymbol{\varepsilon}_n \sim N(0, \sigma_n^2 \mathbf{I}), n = 1, 2, \ldots, N$ and $\xi_s \sim N(0, \tau_s^2 \mathbf{I}), s = 1, 2, \ldots, S$. Hence, using the properties of the probability theory and Bayes' theorem, both the posterior distributions of the latent variables and the measurements can be estimated.

The model parameters of MRPCR is trained using the Expectation-Maximum algorithms. In the E-step, the posterior distributions of the latent variables are estimated using the current parameters as

$$\hat{\boldsymbol{t}}_k = \boldsymbol{\Sigma}_k^{-1}\left(\sum_{n=1}^{N} \sigma_n^{-2} \boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{x}_{nk} + \sum_{s=1}^{S} \tau_s^{-2} \boldsymbol{\Psi}_s^{\mathrm{T}} \boldsymbol{y}_{sk}\right), \tag{2}$$

$$\boldsymbol{\Sigma}_k = \sum_{n=1}^{N} \phi_{nk} \sigma_n^{-2} \boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{\Phi}_n + \sum_{s=1}^{S} \gamma_{sk} \tau_s^{-2} \boldsymbol{\Psi}_s^{\mathrm{T}} \boldsymbol{\Psi}_s + \mathbf{I}, \tag{3}$$

$$\mathrm{E}(\boldsymbol{t}_k \boldsymbol{t}_k^{\mathrm{T}} \,|\boldsymbol{X}_k^o, \boldsymbol{Y}_k^o) = \boldsymbol{\Sigma}_k^{-1} + \hat{\boldsymbol{t}}_k \hat{\boldsymbol{t}}_k^{\mathrm{T}}, \tag{4}$$

where $\hat{\boldsymbol{t}}_k = \mathrm{E}(\boldsymbol{t}_k \,|\boldsymbol{X}_k^o, \boldsymbol{Y}_k^o)$ is the estimated latent variables at the sampling time k and it is based on the current observations, which are noted as $\boldsymbol{X}_k^o$ and $\boldsymbol{Y}_k^o$. In a multi-rate system, the kinds of observations will change over sampling time. Hence, two aided parameters ϕ_{nk} and γ_{sk} are used to represent the collection state of the measurements. If ϕ_{nk} equals to 1, it indicates that the process variables with nth sampling rate have been collected at sampling time k. For contrary, ϕ_{nk} will be zero when these process variables are not available at this time. Such definition of γ_{sk} is also similar. It can be also seen that the model parameters will change over time and they are chosen from $\{\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_N\}$ and $\{\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \ldots, \boldsymbol{\Psi}_S\}$, which is based on how many kinds of process and quality variables are collected at each sampling time. Hence, the proposed MRPCR can be thought as a kind of specific time-variant model.

In the M-step, the model parameters are updated for maximizing the likelihood using the results in the E-step. After taking the first-order derivation of the likelihood, they are given as

$$\hat{\boldsymbol{\Phi}}_n = \left[\sum_{k=1}^{K_n} \boldsymbol{x}_{nk} \hat{\boldsymbol{t}}_k^{\mathrm{T}}\right] \left[\sum_{k=1}^{K_n} \mathrm{E}(\boldsymbol{t}_k \boldsymbol{t}_k^{\mathrm{T}} \,|\boldsymbol{X}_k^o,\ \boldsymbol{Y}_k^o)\right]^{-1}, \tag{5}$$

$$\hat{\boldsymbol{\Psi}}_s = \left[\sum_{k=1}^{J_s} \boldsymbol{y}_{sk} \hat{\boldsymbol{t}}_k^{\mathrm{T}}\right] \left[\sum_{k=1}^{J_s} \mathrm{E}(\boldsymbol{t}_k \boldsymbol{t}_k^{\mathrm{T}} \,|\boldsymbol{X}_k^o,\ \boldsymbol{Y}_k^o)\right]^{-1}, \tag{6}$$

$$\hat{\sigma}_n^2 = \frac{1}{K_n M_n} \sum_{k=1}^{K_n} \left\{ \begin{array}{l} \boldsymbol{x}_{nk}^{\mathrm{T}} \boldsymbol{x}_{nk} - 2\hat{\boldsymbol{t}}_k^{\mathrm{T}} \hat{\boldsymbol{\Phi}}_n^{\mathrm{T}} \boldsymbol{x}_{nk} \\ + \mathrm{tr}(\mathrm{E}(\boldsymbol{t}_k \boldsymbol{t}_k^{\mathrm{T}} \,|\boldsymbol{X}_k^o, \boldsymbol{Y}_k^o)) \hat{\boldsymbol{\Phi}}_n^{\mathrm{T}} \hat{\boldsymbol{\Phi}}_n) \end{array} \right\}, \tag{7}$$

$$\hat{\tau}_s^2 = \frac{1}{J_s H_s} \sum_{k=1}^{J_s} \left\{ \begin{array}{l} \boldsymbol{y}_{sk}^{\mathrm{T}} \boldsymbol{y}_{sk} - 2\hat{\boldsymbol{t}}_k^{\mathrm{T}} \hat{\boldsymbol{\Psi}}_s^{\mathrm{T}} \boldsymbol{y}_{sk} \\ + \mathrm{tr}(\mathrm{E}(\boldsymbol{t}_k \boldsymbol{t}_k^{\mathrm{T}} \,|\boldsymbol{X}_k^o, \boldsymbol{Y}_k^o)) \hat{\boldsymbol{\Psi}}_s^{\mathrm{T}} \hat{\boldsymbol{\Psi}}_s) \end{array} \right\}, \tag{8}$$

where $\mathrm{tr}(\cdot)$ is the matrix trace calculation operator.

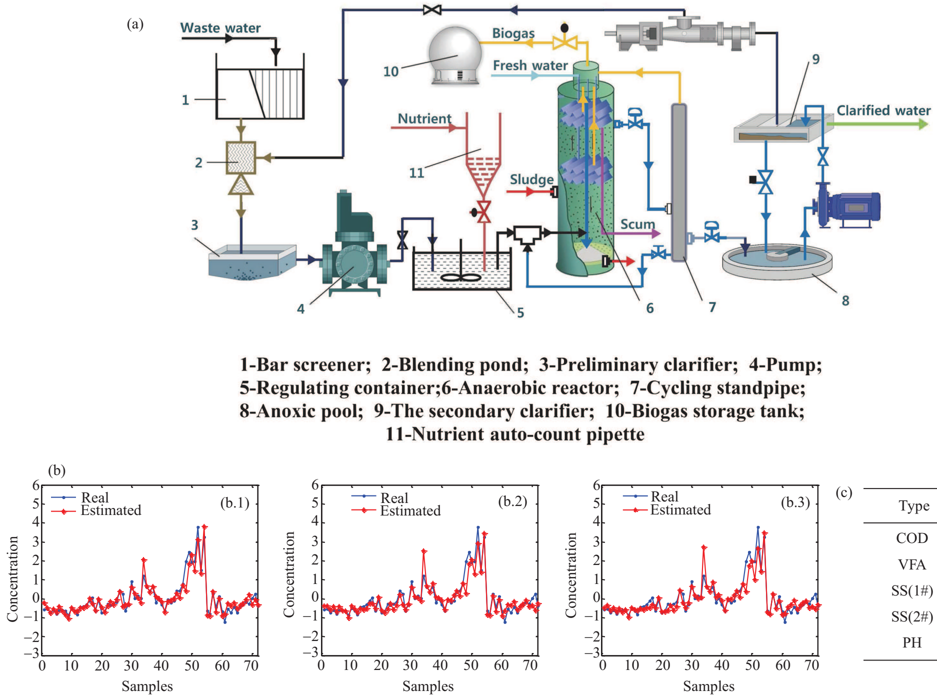
After the model training is finished, the proposed MRPCR can be used for soft sensor purpose. When the test process variable data $\boldsymbol{x}_{n,\mathrm{test}}$ are collected, the quality variables are estimated as

$$\begin{aligned} \boldsymbol{t}_{\mathrm{test}} = & \left(\sum_{n=1}^{N} \phi_{n,\mathrm{test}} \sigma_n^{-2} \boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{\Phi}_n + \mathbf{I}\right)^{-1} \\ & \times \sum_{n=1}^{N} \sigma_n^{-2} \boldsymbol{\Phi}_n^{\mathrm{T}} \boldsymbol{x}_{n,\mathrm{test}}, \end{aligned} \tag{9}$$

$$\boldsymbol{y}_{s,\mathrm{test}} = \boldsymbol{\Psi}_s \boldsymbol{t}_{\mathrm{test}}, \tag{10}$$

where $\phi_{n,\mathrm{test}}$ is the aided parameter to reveal the process variables with the corresponding nth sampling rate are collected or not. Using (10), the quality variables with different sampling rates can be estimated. To evaluate the performance of prediction results, the root mean square error (RMSE) index is used [8].

Simulations. In this section, the proposed MRPCR is demonstrated by a real R2S anaerobic reactor in the papermaking wastewater treatment process. Usually, the main units in a wastewater process contains blending pond, primary clarifier, regulating container, anaerobic reactor, anoxic pool, etc, which is shown in Figure 1(a). Among them, the R2S anaerobic reactor is a crucial unit. In this study, 22 typical variables are collected as the input data. Among them, there are three kinds of sampling rates. Three of them are sampled per hour and 12 of them are sampled per 2 h using the online sensors. Besides, 7 off-line variables in the reactor inlet are also chosen as the input data, which are collected per 24 h. For output data, five major quality indicators are chosen, which contain chemical Oxygen demand (COD), volatile


(a)
Waste water
Biogas
Fresh water
Nutrient
Sludge
Scum
Clarified water
1
2
3
4
5
6
7
8
9
10
11
1-Bar screener; 2-Blending pond; 3-Preliminary clarifier; 4-Pump;
5-Regulating container;6-Anaerobic reactor; 7-Cycling standpipe;
8-Anoxic pool; 9-The secondary clarifier; 10-Biogas storage tank;
11-Nutrient auto-count pipette
(b)
Real
Estimated
(b.1)
(b.2)
(b.3)
Concentration
Samples
6
5
4
3
2
1
0
−1
−2
−3
0
10
20
30
40
50
60
70

(c)

Type	MRPCR	PPCR	PLS
COD	0.6730	0.7167	0.7295
VFA	0.6685	0.6898	0.6999
SS(1#)	0.3530	0.4093	0.4091
SS(2#)	0.3249	0.3641	0.3890
PH	0.6650	0.7042	0.6956

Figure 1 (Color online) (a) The flowchart of the papermaking wastewater treatment process; (b) the prediction results of the suspended solids (2#) in the anaerobic reactor outlet using (b.1) MRPCR, (b.2) PPCR, and (b.3) PLS; (c) the prediction results in anaerobic reactor outlet.

fatty acid (VFA), PH, suspended solids (SS) 1# and 2# in the anaerobic reactor outlet. All of these quality variables are tested at the lab and collected per 24 h. Using a period of three months data in the normal condition, the proposed MRPCR model is constructed. For comparison, two traditional soft sensor model PPCR and PLS are also built. To validate the prediction performance of the proposed method, another period of 72 days data in the same unit are collected. The prediction results are given in Figure 1(c). It can be seen that the RMSE using MRPCR for all the five quality data performs better than the alternatives. The detailed prediction results for SS(1#) are given in Figure 1(b). The main reason is that the proposed method has utilized all the measurements instead of dropping some useful information or making the redundant estimation for down-sampling or upsampling.

Conclusion and future work. In this study, a multi-rate PCR model is derived to integrate multi-rate measurements for quality prediction purpose. Using MRPCR, the prediction ability is improved since all the process and quality data have been utilized without own-sampling or upsampling.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61603342), NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (Grant No. U1609214), and China Postdoctoral Science Foundation (Grant No. 2018M630674)

References

1 Ge Z, Song Z, Ding S X, et al. Data mining and analytics in the process industry: the role of machine learning. IEEE Access, 2017, 5: 20590–20616

2 Zhou L, Zheng J, Ge Z, et al. Multimode process monitoring based on switching autoregressive dynamic latent variable model. IEEE Trans Ind Electron, 2018, 65: 8184–8194

3 Yuan X, Wang Y, Yang C, et al. Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes. IEEE Trans Ind Electron, 2018, 65: 1508–1517

4 Ge Z. Process data analytics via probabilistic latent variable models: a tutorial review. Ind Eng Chem Res, 2018, 57: 12646–12661

5 Zhou L, Chen J, Song Z, et al. Semi-supervised PLVR models for process monitoring with unequal sample sizes of process variables and quality variables. J Process Control, 2015, 26: 1–16

6 Yao L, Ge Z. Locally weighted prediction methods for latent factor analysis with supervised and semisupervised process data. IEEE Trans Automat Sci Eng, 2017, 14: 126–138

7 Shao W, Tian X. Semi-supervised selective ensemble learning based on distance to model for nonlinear soft sensor development. Neurocomputing, 2017, 222: 91–104

8 Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)? — Arguments against avoiding RMSE in the literature. Geosci Model Dev, 2014, 7: 1247–1250