Probabilistic Generative Model with Long-Term Memory and its Application in Chemical Process Modeling

Zhichao Chen[1]， Luyao Wang[1], Guofei Chen[1], Zhiqiang Ge[1]

（[1]State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, China）

Abstract

Probabilistic generative model (PGM) is being widely concerned in data-driven based chemical process modeling filed for its capability in the giving the value and distribution of quality variables. However, traditional PGM based on deep learning could not extract the feature in temporal scale well, which results in the block of the application in chemical process. To solve this problem, a novel model named probabilistic generative model with long-term memory (LTM-PGM) is proposed in this work. Comparing to existing models, the LTM-PGM is capable of extracting time-variant feature and decreasing the modeling parameters simultaneously, by which the nonlinearity and dynamic of the chemical process can be modeled. Finally, the proposed model is verified in an industrial process to show its effectiveness and superiority.

Key words： Probabilistic Generative Model, Deep Learning, Chemical Process Modeling, Dynamic Data Modeling

Introduction

In chemical process, the quality indicators that are difficult to measure online play an important role in process monitoring, control, and optimization[1]. To measure the quality variables in real-time, the regression models can be adopted by bridging connections between quality variables and process variables that are available online. As the chemical processes are too complex to obtain prior knowledge, the data-driven models are being well developed for quality variables prediction in the past few decades. Traditional machine learning methods like principal component regression (PCR)[2], partial least square regression (PLSR)[3], artificial neural network (ANN)[4], and supporting vector regression (SVR)[5] has been widely used. Different from the deterministic model, the probabilistic models have been paid well attention in recent years. The stochastic models especially the probabilistic generative model (PGM) obtain the quality variables in distribution form, which has a good robustness when facing the process noise as well as the capability of handling missing value.

On the other hand, the development of deep learning (DL) technology[6] is gradually used in chemical process modeling for the neural network with enough layer and hidden nodes can model the nonlinearity and dynamic of process. Conventional deep learning models like (stacked) auto encoder (SAE)[7], restricted Boltzmann machine (RBM)[8], recurrent neural network (RNN)[9], and convolution neural network (CNN)[10] have been adopted in the quality variables prediction.

Despite the probabilistic generative model and deep learning model have attained great success in chemical process, the combination of the abstract feature extraction capability of the DL model and the robustness of the process noise of the PGM model remains a challenge. One of the representative models that combines the PGM and DL is variational auto encoder[11] (VAE). Based on the VAE model, Xie et al[12]. analyzed the similarity of the latent space of the supervised VAE and unsupervised VAE. By transferring the latent space from the unsupervised one to the supervised one, the soft-sensor model is constructed. Despite the model can handle the missing value, the dynamic features of the process

are ignored. Meanwhile, based on the hierarchy VAE soft-sensor[13], Shen et al. make a first-order Markov assumption on the latent space in time sequential. The nonlinear dynamic system (NDS) model assisted by VAE is constructed[14] and the variable correction in spatial scale and temporal scale are proceeded in the latter work[15]. However, in the modeling process, to avoid the trivial solution of the AE, the dimension of the latent space will be less than that of the process variables, which higher-order dynamic feature will be ignored. Hence, the dynamic relationship of the process variables may not be well considered.

Considering the dynamic of process variables, the probabilistic generative model with long-term memory (LTM-PGM) is proposed in this work. In the model, not only the NDS is considered for the sake of extracting the features of process variables in short-term, but also the gated recurrent unit (GRU) is designed to memorize the long-term information from the process variables. The nonlinear feature extraction of VAE and the special structure of the model can enhance the model in the nonlinearity and dynamic modeling capability. The innovation of the work can be summarized as follow:

1). The long-term feature extraction is first considered in the data-driven chemical process soft sensor model based on PGM

2). By analyzing the evidence lower bound optimization (ELBO), the loss function that fits the stochastic gradient descent (SGD) framework is derived strictly of the LTM-PGM model.

The rest of the paper is organized as follows: in section 1, the principle of the PGM model is reviewed. In section 2, the LTM-PGM model is formulated and the corresponding loss function is given. After that, an industrial debutanizer column is performed in section 3 to show the effectiveness and superiority of the LTM-PGM model comparing to other deep learning models and the conclusions are drawn in the section 4.

1 Review of the PGM model

1.1 The principle of VAE and NDS

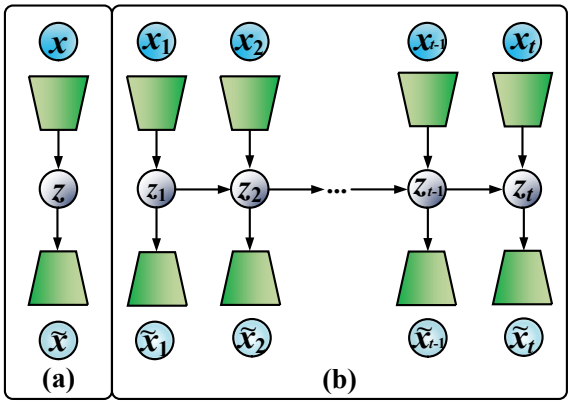

Fig. 1 The model structure (a). VAE, (b). NDS

Fig. 1 (a) presents the structure of VAE. The VAE consists of the inference network (encoder) and generative network (decoder). In the inference network, the variable x is mapped by the encoder into stochastic variable z in the latent space. After that, the stochastic variable z is mapped via the generative network in the decoder to reconstruct variable x. The objective function of the VAE is to maximize the log-likelihood function of the marginal probability of variable x. By the inequality shown in Eq. (1), the objective function can be transformed into the ELBO for the non-negative property of the Kullback-Leiber divergence.

$$\begin{aligned}\log p(x) &= E_{q_\phi(z|x)}[\log p_\theta(x)] \\ &= E_{q_\phi(z|x)}[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}] + E_{q_\phi(z)}[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}] \\ &= ELBO + KL(q_\phi(z|x)\|p_\theta(z|x)) \geq ELBO\end{aligned} \tag{1}$$

where the $q(z)$ is the variational distribution family, Φ and θ are the parameters of the encoder and decoder, respectively.

The model of the linear dynamic system (LDS)[16] is given in Eq. (2). Based on LDS and VAE, if the z and x are calculated via encoder and decoder respectively, the NDS presented in Fig. 1 (b) can be obtained.

$$\begin{cases} z_n = Az_{n-1} + w_n, W \sim \mathsf{N}(0,\Gamma) \\ x_n = Cz_{n-1} + v_n, u \sim \mathsf{N}(0,\Sigma) \\ z_1 = \mu_0 + u, u \sim \mathsf{N}(0,V_0) \end{cases} \tag{2}$$

According to Fig. 1 (b), the joint probability of

the x and z can be decomposed as Eq. (3), while the variational distribution can be decomposed as Eq. (4) shown.

$$p(x,z) = p(z_1)p(x_1|z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})p(x_t|z_t) \tag{3}$$

$$q_\phi(z) = q_{\phi_1}(z_1)\prod_{t=2}^{T} q_{\phi_t}(z_t|z_{t-1}) \tag{4}$$

Substitute Eq. (3) and (4) into the ELBO given in Eq. (1), and the ELBO for the NDS can be rewritten as Eq. (5) shown. Note that, the subscript t indicates that the parameters of the inferential network and the generative network will change in temporal scale.

$$\begin{aligned} ELBO &= E_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p(z)) \\ &= -KL(q_{\phi_1}(z_1|x)\|p(z_1)) + E_{q_{\phi_1}(z_1|x_1)}[\log p_{\theta_1}(x_1|z_1)] \\ &\quad -\sum_{t=2}^{T} KL(q_{\phi_t}(z_t|z_{t-1})\|p(z_t|z_{t-1})) \\ &\quad +\sum_{t=2}^{T} E_{q_{\phi_t}(z_t|z_{t-1})}[\log p_{\theta_t}(x_t|z_t)] \end{aligned} \tag{5}$$

1.2 The principle of PGM modeling for soft-sensor

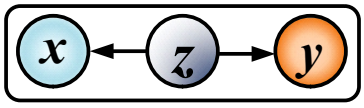

Fig. 2 The graphical structure of the PGM model

When the label is introduced in the generative network, the graphical structure of the decoder can be presented in Fig. 2. By analyzing the graphical structure according to the Bayesian network, the joint probabilistic distribution density function can be represented as Eq. (6).

$$p(x,y,z) = p(z)p(x,y|z) \tag{6}$$

If the z is known, the relationship of the x and y written as $x \perp\!\!\!\perp y|z$ set up. Therefore, in the generative network, the expression can be formulated as Eq. (7).

$$p(x,y|z) = p(x|z)p(y|z) \tag{7}$$

Thus, the ELBO of the VAE can be formulated into supervised form as Eq. (8) shown. As a result, the model can be used for chemical process modelling.

$$\begin{aligned} ELBO &= E_{q_\phi(z|x)}[\log p_\theta(x,y|z)] - KL(q_\phi(z|x)\|p(z)) \\ &= E_{q_\phi(z|x)}[\log p_{\theta_x}(x|z)] + E_{q_\phi(z|x)}[\log p_{\theta_y}(y|z)] \\ &\quad -KL(q_\phi(z|x)\|p(z)) \end{aligned} \tag{8}$$

2 Model Formulation of LTM-PGM

The Markov property of the LDS restrict long term feature learning of NDS. To solve this problem, the LTM-PGM model shown in Fig. 3 is proposed. By using the GRU cell to memorize the time varying feature of the process (process the input x at each time t same as the linear state space model), the model can extract long term feature. The detailed model will be formulated in this section.

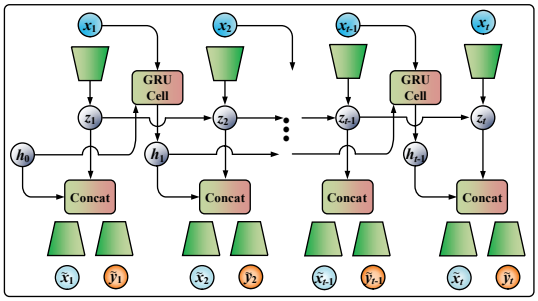

Fig. 3 The model structure of the LTM-PGM

2.1 The expression of the GRU cell

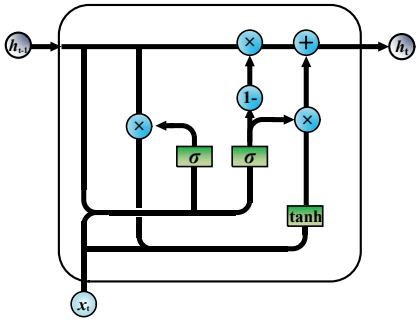

Fig. 4 The model structure of the GRU

Fig. 4 presents the structure of the GRU cell[17], to memorize the long-term feature of the process variables, the expressions of the reset gate and update gate are given in Eq. (9) and (10).

$$u_t = \sigma(W_u[x_t, h_{t-1}] + b_u) \tag{9}$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \tag{10}$$

By calculating the gate control signal, the memory of last time can be reset according to Eq. (11).

$$\tilde{h}_t = \sigma(W_h[x_t, (r_t \quad h_{t-1})] + b_h) \tag{11}$$

The reset memory can be update via Eq. (12), by which the memory of this time can be obtained, which means that the dynamic feature less than time t can be obtained (denoted by subscript $\leq$). By using same parameters to extract the feature of the input sequence recursively, the feature of the secondary variables can be obtained.

$$x_{\leq t} \quad h_t = u_t \quad h_{t-1} + (1 - u_t) \quad \tilde{h}_t \tag{12}$$

2.2 The derivation of the loss function

As the long-term memory is considered in generative network, the expression of posterior distribution should consider the information of the process variables less than time t. Therefore, the joint probability density function can be decomposed as Eq. (13) shown. Correspondingly, the variational distribution can be decomposed into Eq. (14).

$$p(x, y, z) = p(z_1)p(x_1, y_1 | z_1)\prod_{t=2}^{T} p(z_t | z_{t-1})p(x_t, y_t | z_t, x_{<t}) \tag{13}$$

$$q_\phi(z) = q_\phi(z_1)\prod_{t=2}^{T} q_\phi(z_t | z_{t-1}) \tag{14}$$

Substituting the Eq. (13) and Eq. (14) into Eq. (3), the ELBO can be reformulated as Eq. (15) shown.

$$\begin{aligned} ELBO = & -KL(q_\phi(z_1|x_1) \| p(z_1)) + E_{q_\phi(z_1|x_1)}[\log p_\theta(x_1, y_1 | z_1)] \\ & -\sum_{t=2}^{T} KL(q_\phi(z_t | z_{t-1}) \| p(z_t | z_{t-1})) \\ & +\sum_{t=2}^{T} E_{q_\phi(z_t|z_{t-1})}[\log p_\theta(x_t, y_t | z_t, x_{<t})] \end{aligned} \tag{15}$$

Note that, under the SGD framework, the loss function can be converted as Eq. (16) shown.

$$\max \ ELBO \Leftrightarrow \min \ Loss \quad -ELBO \tag{16}$$

Thus, the prior of the latent variables are given as Eq. (17) shown. After re-parameterization trick, the loss function of the model can be given as Eq. (18) shown.

$$\begin{cases} p(z_1) \sim \mathrm{N}(0, I) \\ p(z_t | z_{t-1}) \sim \mathrm{N}(0, I) \end{cases} \tag{17}$$

$$\begin{aligned} & Loss \\ & = \frac{1}{T \times N}\sum_{t=1}^{T}(\|x_t - \tilde{x}_t\|_F^2 + \|y_t - \tilde{y}_t\|_F^2) + KL(q_\phi(z_1|x_1) \| p(z_1)) \\ & +\sum_{t=2}^{T} KL(q_\phi(z_t | z_{t-1}) \| p(z_t | z_{t-1})) \end{aligned} \tag{18}$$

2.3 The quantity of the model parameters

In this section, considering the encoder and the decoder of the NDS and LTM-PGM are in the same structure, the number of the parameters in the latent space will be analyzed to demonstrate the superiority of the LTM-PGM. In the NDS, the number of the parameters in the latent space is given in Eq. (19).

$$N_{NDS} = TN_z^2 + TN_z^2 = 2TN_z^2 \tag{19}$$

Thanks to the introduction of GRU, the time varying feature can be memorized and the coefficient of the LDS part in LTM-PGM can remains unchanged as the time goes. Thus, the quantity of the model parameters is given as Eq. (20).

$$N_{NDS} = N_z^2 + 3N_xN_z + 3N_h^2 + 3N_h \tag{20}$$

By comparing the Eq. (19) and (20), the complexity of the model decreases from $O(N^3)$ to $O(N^2)$. As a result, the training cost of the model will decrease when the sequence length of the training data is long.

2.4 Soft Sensor Development

Table 1 presents the algorithm for soft sensor construction based on LTM-PGM. In the off-line stage, after the model will be trained from the training dataset. After that, in the on-line stage, the model will predict the quality variables when a piece of new data samples from the process measuring

meters.

Table 1. The algorithm for soft sensor construction

Off-line Stage	
1	Determine the model structure and hyper-parameters
2	Obtain the historical data and scale the data
3	Train the LTM-PGM model using SGD
On-line Stage	
4	Obtain the sample
5	Predict the quality indicators using trained LTM-PGM

To evaluate the effectiveness of the model, the root mean square error (RMSE) and correlation coefficient (R^2) defined in Eq. (21) and (22) are adopted.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \tilde{y}_n)^2}{N}} \tag{21}$$

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \tilde{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2} \tag{22}$$

where the y_n is the real value and the $\tilde{y}_n$ is the predicted value of the model, n is the total number of the testing dataset. The smaller the RMSE and the bigger the R^2, the better the model performance.

3 Case Study

In this section, the LTM-PGM framework will be evaluated on a real industrial process named debutanizer column. In this section, the process of debutanizer column (DC) is described at first, and then the chemical process model is constructed to evaluate the effectiveness of the LTM-PGM model based on Python 3.8.

3.1 The process description of the DC

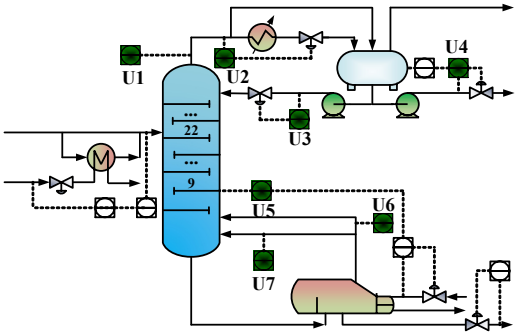

Fig. 5 The flowsheet of the DC

Fig. 5 presents the flowchart of the debutanizer column. Debutanizer column is an important part of the de-sulfuring and naphtha splitter plant in the refinery[18]. As the debutanizer column is required to maximizing the pentane (C5) content in the overheads distillate and minimize the butane (C4) content in the bottom flow simultaneously, to improve the performance of the quality control, it is necessary to carry out the real time prediction of the butane content at the bottom flow. However, the butane content is not directly on the bottom flow, but on the overheads of the sequential deisopentanizer column by the gas chromatograph results in a large measuring delay.

To deal with the problem mentioned above as well as improve the control quality of the column, soft sensors have been adopted to measure the butane concentration from the bottom flow in real time. Several hard sensors are installed on the plant to collect process variables at real time, which can be used for soft sensor secondary variables. Seven variables are chosen for soft sensor modeling, the detailed descriptions are given in Table 2 and the location of the hard sensors are marked in orange circle in Fig. 5. There exist complicated nonlinearities between variables involved in the debutanizer column. Therefore, it is necessary to carry out nonlinear soft sensor modeling for prediction performance improvement. A total amount of 2,394 data has been collected from the process, the first 1,596 data are utilized as the training data, the remaining 798 data are used as the testing data to evaluate the effectiveness of the model.

Table 2. The algorithm for soft sensor construction

Process Variables	Unit	Description
U1	°C	Top temperature
U2	$kg \cdot cm^{-2}$	Top pressure
U3	$m^3 \cdot h^{-1}$	Reflux flowrate
U4	$m^3 \cdot h^{-1}$	Top distillate rate
U5	°C	Temperature of 9th tray
U6	°C	Bottom temperature A
U7	°C	Bottom temperature B

3.2 Model evaluation and comparison

By trail-and-error, the sequence length of the LTM-PGM is set to be 10, and the latent space dimension is set to be 10, the hidden layer dimension of the GRU is set to be 15 to construct the soft sensor model. To show the effectiveness and superiority of the model, the variable-wise weighted stacked autoencoder (VW-SAE), supervised LSTM (S-LSTM)[19], and variable-wise attention LSTM (VA-LSTM) [20] are adopted. The predicted value of the model mentioned above are presented in Fig. 6 and the evaluation results are presented in Table 3.

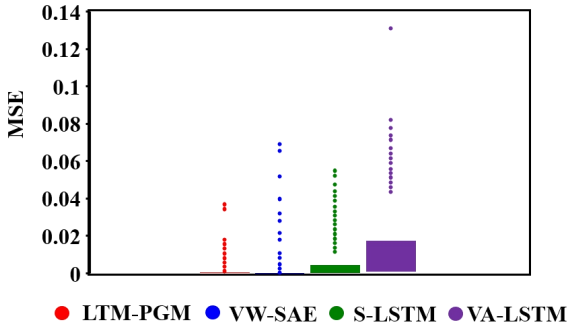

Fig. 6 The box-plot of prediction mean square error predicted value of the models

The box-plot of prediction error on the testing dataset is presented in Fig. 6. The widen gap between the LTM-PGM and other baseline models shown that the existence of KL divergence can help LTM-PGM not only avoid over-fitting phenomenon but also enhance the model expression in process non-linearity. Meanwhile, it should be noted that, the LTM-PGM can extract dynamic feature with the help of the GRU cell and results a better performance comparing to the VW-SAE. In Table 3. The RMSE of the LTM-PGM is 10.88~297.83% lower than that of other models while the R^2 is 0.29~12.75% higher. Therefore, the LTM-PGM can make a more effective prediction comparing to other models according to statistical index.

Table 3. The evaluation results

Model Name	RMSE	R^2
VA-LSTM	0.1115	0.8797
SLSTM	0.0721	0.9853
VW-SAE	0.0311	0.9890
LTM-PGM	**0.0280**	**0.9919**

4 Conclusion

This work proposed the LTM-PGM model and demonstrated its application in chemical process modeling. By using the GRU to capture the dynamic of the process variables, the model can decrease the parameters and extract long term feature simultaneously. Finally, a case study on a real debutanizer was proposed to demonstrate the effectiveness and superiority of the model.

It should be pointed out that during the derivation of the loss function, the long-term memory was merely presented in the generative network. In future works, the concentration will be paid on the representation of the long-term memory in the inferential network and generative network simultaneously.

Notations

b——bias of the neural network

h——hidden state of neural network

$\tilde{h}$ ——long term memory of neural network

r——reset gate

u——update gate

v——noise of the observation value

W——weight of neural network

w——noise of the state transition

x——input value

y——output value

z——hidden state of state space model

GRU——gate recurrent neural network

RMSE——root mean square error

R^2——correlation coefficient

Reference

[1]. Huang B, Kadali R. **Dynamic modeling, predictive control and performance monitoring: a data-driven subspace approach [M].** Springer, 2008.

[2]. Ge Z, Huang B, Song Z, Mixture semi-supervised principal component regression model and soft sensor application [J]. **AIChE Journal**, 2014, 60(2): 533-545.

[3]. Qin S, Recursive pls algorithms for adaptive data modeling [J]. **Computers & Chemical Engineering**, 1998, 22(4): 503–514.

[4]. Gonzaga J, Meleiro L, Kiang C, *et al*. ANN based soft-sensor for real-time process monitoring and control of an industrial polymerization process [J]. **Computers & Chemical Engineering**, 2009, 33 (1): 43-49.

[5]. Kaneko H, Funatsu K. Application of online support vector regression for soft sensors [J]. **AIChE Journal**, 2014, 60(2): 600-612.

[6]. Sun Q, Ge Z. A survey on deep learning for data-driven soft sensors [J]. **IEEE Transactions on Industrial Informatics**, 2021, 17 (9): 5853-5866.

[7]. Yuan X, Huang B, Wang Y, *et al*. Deep learning based feature representation and its application for soft sensor modeling with variable-wise weighted SAE [J]. **IEEE Transactions on Industrial Informatics**, 2018, 14 (7): 3235-3243.

[8]. Shang C, Yang F, Huang D, *et al*. Data-driven soft sensor development based on deep learning technique [J]. **Journal of Process Control**, 2014, 24 (3): 223-233.

[9]. Yuan X, Li L, Shardt Y., *et al*. Deep learning with spatiotemporal attention-based lstm for industrial soft sensor model development [J]. **IEEE Transactions on Industrial Electronics**, 2021, 68(5): 4404-4414.

[10]. Yuan X, Qi S, Wang Y, *et al*. A dynamic cnn for nonlinear dynamic feature learning in soft sensor modeling of industrial process data [J]. **Control Engineering Practice**, 2020, 104: 104614.

[11]. Kingma D P, Welling M. Auto-encoding variational bayes [C]. **Proceedings of the 2nd International Conference on Learning Representations (ICLR)** , Canada, 2014.

[12]. Xie R, Jan N, Hao K, *et al*. Supervised variational autoencoders for soft sensor modeling with missing data [J]. **IEEE Transactions on Industrial Informatics**, 2020, 16 (4): 2820-2828.

[13]. Shen B, Yao L,Ge Z. Nonlinear probabilistic latent variable regression models for soft sensor application: From shallow to deep structure [J]. **Control Engineering Practice**, 2020, 94: 104198.

[14]. Shen B, Ge Z. Supervised nonlinear dynamic system for soft sensor application aided by variational auto-encoder [J]. **IEEE Transactions On Instrumentation and Measurement**, 2020, 69 (9): 6132-6142.

[15]. Shen B, Ge Z. Weighted nonlinear dynamic system for deep extraction of nonlinear dynamic latent variables and industrial application [J]. **IEEE Transactions on Industrial Informatics**, 2021, 17 (5): 3090-3098.

[16]. Bishop C. M., **Pattern Recognition and Machine Learning**[M]. London, U.K.: Springer, 2006.

[17]. Goodfellow I, Bengio Y, Courville A. **Deep learning**[M]. MIT press, 2016.

[18]. Fortuna L, Graziani S, Xibilia M G. Soft sensors for product quality monitoring in debutanizer distillation columns [J]. **Control Engineering Practice**, 2005, 13(4): 499-508.

[19] Yuan X, Li L, Wang Y. Nonlinear dynamic soft sensor modeling with supervised long short-term memory network[J]. **IEEE Transactions on Industrial Informatics**, 2019, 16(5): 3168-3176.

[20] Yuan X, Li L, Wang Y, *et al*. Deep learning for quality prediction of nonlinear dynamic processes with variable attention‐based long short‐term memory network [J]. **The Canadian Journal of Chemical Engineering**, 2020, 98(6): 1377-1389.