Reliable Data-Driven Soft Sensor Modeling with the Aid of Stable Loss Function and Sample Graph

Ruikun Zhai,a Jinchuan Qian,a Xiaoyu Jiang,a Xinmin Zhang,a,\* Zhihuan Song,a Manabu Kanob

aThe State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China

bDepartment of Systems Science, Kyoto University, Kyoto 606-8501, Japan

xinminzhang@zju.edu.cn

Abstract

Data-driven soft sensors, as a replacement or complement of physical sensors, have been prevalent in predicting hard-to-measure key quality indicators of industrial chemical processes. However, with the existence of noise, outliers, and process drifts in process data, data-driven soft sensors suffer from poor reliability. Based on the manifold regularization framework, this paper proposes two new soft sensor models, namely Laplacian Huber regression and Laplacian piecewise-linear regression, in which new loss functions that are more stable in the face of noise and outliers are proposed, and theprior knowledge of the target process is injected into the learning objectives in the form of a graph Laplacian. The graph Laplacian is derived from an undirected graph obtained by combining prior knowledge and historical data, which embeds the relationships between process samples and helps form an intrinsic regularization term. The new loss functions together with the intrinsic regularizer can guide the learning process of the model in the correct direction, keeping the model from overfitting and being misled by noise or outliers. New optimization methods have also been developed to efficiently solve the learning objectives of the proposed models. The improved reliability performance of the newly proposed models has been validated by a simulation study and a case study of the real-world high-low transformer unit process.

**Keywords**: Data-driven Modelling, Reliable Soft Sensors, Manifold Regularization, Graph Laplacian

* 1. Introduction

In modern industry, fast and accurate measurement and analysis of key quality indicators (KQIs) are of great significance to improve process safety and product quality (Kano and Fujiwara, 2013). Unfortunately, many of the process KQIs are impossible, difficult, or costly to measure in real time (Luo et al., 2023). To tackle this problem, soft sensing technology has been rapidly developed in the last decades (Zhang et al., 2023). To build a soft sensor, either a first principle-based approach or a data-driven approach can be typically used (Sun and Ge, 2021). As modern processes become increasingly complicated, it becomes increasingly difficult to build accurate first principle-based models. On the other hand, with the development of distributed control systems and computer technology, data-driven soft sensor modeling is becoming increasingly popular. However, data-driven models are prone to overfitting or are sensitive to noise and outliers. Therefore, it is very meaningful to build reliable soft sensor models.

To improve the reliability of data-driven models, there are some studies in the literature that inject prior knowledge of the process into data-driven modeling process (Westerhuis et al., 2007). They sought to add knowledge-specific regularization terms to the learning objective of data-driven models. As a common form of knowledge, graph structures can be found in many aspects of the process industry, e.g., process diagrams. Such knowledge can be abstracted into undirected graphs, whose graph Laplacian can then be used to form a regularization term. In the manifold regularization framework (Belkin et al., 2006), the graph Laplacian of training samples served as a discrete approximation of the Laplace-Beltrami operator on the data manifold. The quadratic form of model predictions and the graph Laplacian (called the *intrinsic* regularization term) is added to the learning objective of the function estimation framework in the reproducing kernel Hilbert space (RKHS) (Pillonetto *et al.*, 2022) and the Laplacian support vector regression (LapSVR) model is proposed. Some models have also been proposed based on the manifold regularization framework, e.g., semi-supervised hierarchical extreme learning machine (Yao and Ge, 2018). However, most contributions under the manifold regularization framework have been focused on the semi-supervised setting. Nevertheless, it is argued that manifold regularization can also prevent data-driven models from overfitting and being affected by data noise and outliers, which is an enjoyable merit for both supervised and semi-supervised learning scenarios.

To further improve the reliability of soft sensors, this study proposes two new reliable soft sensor models under the manifold regularization framework, namely Laplacian Huber regression (LapHBR) and Laplacian pointwise-linear regression (LapPLR). In LapHBR and LapPLR, efficient optimization methods are proposed to solve the learning objectives of the two models. In addition, a general approach for extracting graph information based on the historical time series of process variables is proposed, which is a method to obtain the graph Laplacian operator in the model. The effectiveness of the proposed LapHBR and LapPLR models was verified through a simulation experiment and a real-world industrial process. Application results demonstrate that the proposed two models exhibit higher reliability compared to other conventional soft sensor models.

* 1. Reliable Soft Sensing in Manifold Regularization Framework
     1. Graph Laplacian

Assume that the curve of the KQI variable with time is smooth. Then, construct the graph with vertices, and then perform the following operations on all the samples to construct the weight matrix :

* Given a sample and a pre-determined positive integer , only the samples whose timestamps fall within the time window (excluding sample ) may have an edge with sample .
* Only the samples (, and in this paper is set to 5) inside the window that have the smallest Euclidean distances from sample are connected to the -th sample with the connection weight .

Update by to guarantee that the graph is undirected. Finally, the graph Laplacian can be obtained by , where is diagonal with diagonal entries .

The proposed approach is a hybrid approach combining prior knowledge and process data. It is also a general approach that can be applied to all time-series modeling scenarios.

* + 1. Manifold Regularization Framework

Assume that each sample has process variables, which are drawn from the marginal distribution , where is the underlying data manifold. Further, it is assumed that conditional distributions and are similar, if two data points and are close in . Belkin et al. (2006) claimed that the graph can serve as a discrete approximation of , and proposed the manifold regularizer , where , resulting in the *manifold regularization framework*:

|  |  |
| --- | --- |
|  | (1) |

where is the coefficients of the function in RKHS induced by a Mercer kernel with , and is the kernel matrix whose -th entry is .

* + 1. Reliable Soft Sensor Models: LapHBR and LapPLR
       1. Laplacian Huber Regression

The Huber loss is defined as

|  |  |
| --- | --- |
|  | (2) |

where is a hyper-parameter and is the error. The advantage of Huber loss is that it is less sensitive to outliers, since it grows linearly for errors . Therefore, this study proposes LapHBR, where the data loss *l* in Eq. (1) is instantiated as the Huber loss in Eq. (2), and an additional bias parameter is added. With the help of Lagrange multipliers , we can calculate the dual problem of the proposed LapHBR model as follows:

|  |  |
| --- | --- |
|  | (3) |

This dual problem is a quadratic programming (QP) problem with linear and box constraints that can be solved more efficiently than the primal problem by convex optimization algorithms, such as the interior-point methods (Boyd and Vandenberghe, 2004). After solving the dual problem, the values of and can be obtained in sequence and thereby obtaining the optimal LapHBR model.

* + - 1. Laplacian Piecewise-Linear Regression

The LapSVR model under the manifold regularization framework uses the -insensitive loss, which completely ignores errors less than or equal to , and consequently may cause the model to underfit. On the other hand, the Huber loss uses the squared error on small , which may cause the model to overfit. Based on the above thoughts, a new loss called pointwise-linear (PL) loss is proposed as follows:

|  |  |
| --- | --- |
|  | (4) |

where . The PL loss also takes the small errors into account, and overfitting and underfitting w.r.t. small errors can be balanced by tuning the parameter. Note that the -insensitive loss and the absolute error loss are two extreme cases of the newly proposed PL loss. Based on this, the LapPLR model is proposed, whose objective function is Eq. (1) with the loss function instantiated by the PL loss in Eq. (4) and bias .

Likewise, the dual problem of LapPLR can be obtained,

|  |  |
| --- | --- |
|  | (5) |

where are slack variables. This dual form is also a QP problem with linear equality and linear inequality constraints that can be solved efficiently by QP solvers.

* 1. Case Study
     1. Simulation Example

The simulation experiment was implemented based on 3 components, , and . The data were generated by

|  |  |
| --- | --- |
|  | (6) |

where is random additive noise and represents the outlier component in the generation of . Given a predefined hyperparameter , have a probability of being 10 (the outlier value), and probability of being 0.

Different sets of data were generated with different values of , ranging from 0 % to 40 %. Each set of the data comprises 2000 training data points and 2000 test ones. The mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination () were used to evaluate the models.

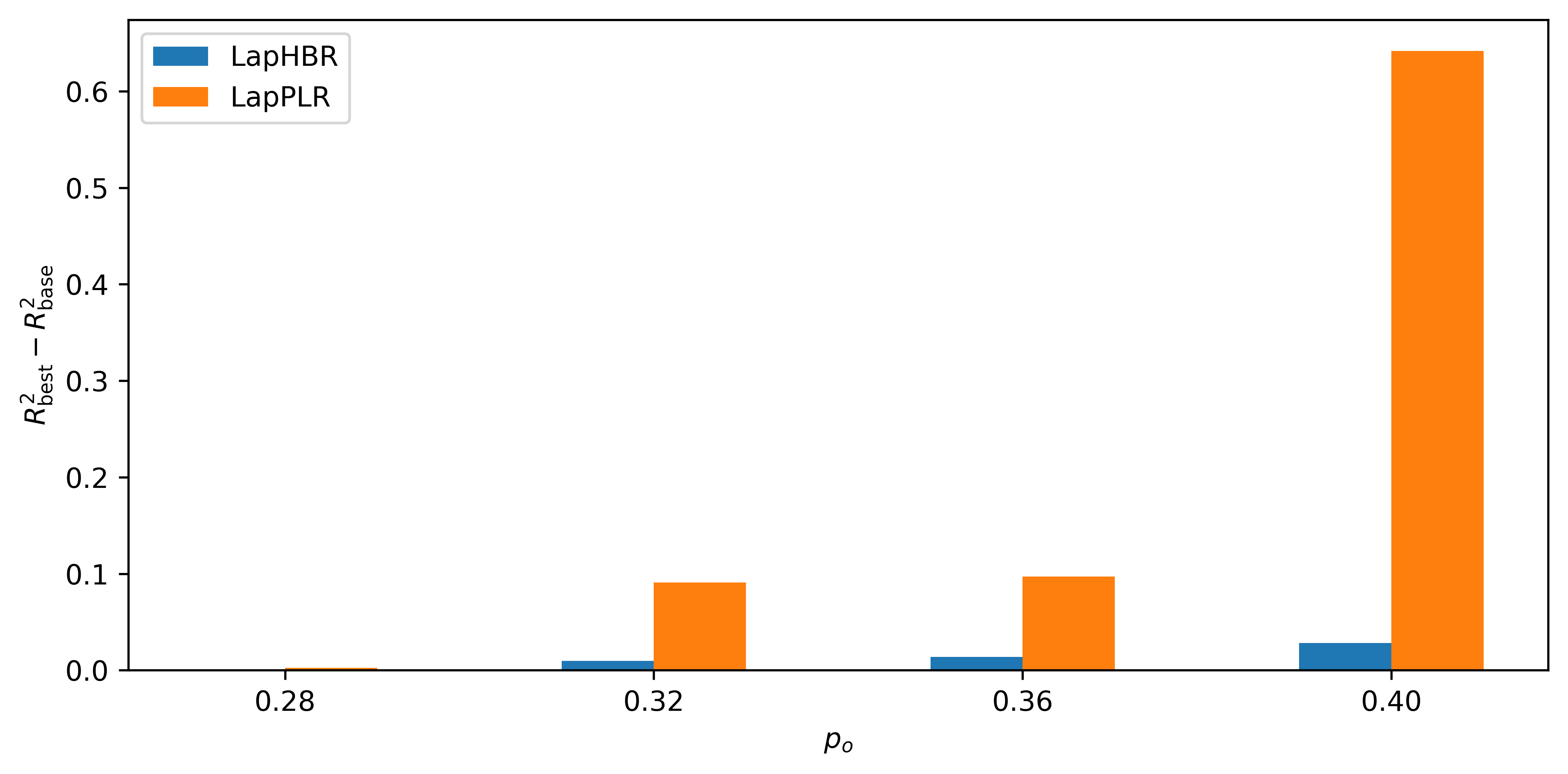


Figure 1. Performance of different models under different values. Left: the test of different models. Right: The relative improvements of LapHBR and LapPLR, where is the value with and is the value with .

When , the best performances of LR, MLP, LapPLR, LapHBR, and LapSVR with different values are shown in Table 1.

Table 1. Best performance of the models in comparison under outlier ratio . The second column indicates the value of the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model |  | Test | Test MSE | Test MAE |
| LR | N/A | -7.5650 | 5.8144 | 2.3641 |
| MLP | N/A | -8.6069 | 6.5217 | 2.4342 |
| LapSVR | 0.01 | 0.7334 | 0.1810 | 0.3694 |
| **LapHBR** | **0.01** | **0.9070** | **0.0632** | **0.1968** |
| **LapPLR** | **1.0** | **0.9385** | **0.0418** | **0.1330** |

The MLP and LR fail to capture the real relationship between the process variables and the KQI variable. The LapSVR model struggled to give moderate-level predictions, but its performance is far behind those of LapHBR and LapPLR. In this scenario, the poor performance of LR and MLP results from their squared loss function, which focuses too much on the small losses, thus being overly affected by the outliers. Moreover, all three models LapSVR, LapHBR and LapPLR achieved the best performance with a positive , which implies that the incorporation of the proposed graph Laplacian can also improve the reliability of soft sensor models.

The test of the LR, LapSVR, LapHBR and LapPLR models under different values of is shown in the left part of Figure 1. As grows, the performance of LapSVR starts to drop. Contrastively, LapHBR and LapPLR perform stably with the increase of outliers.

The effect of *a priori* sample graph is illustrated by recording the relative improvements of the test of LapPLR and LapHBR when the graph Laplacian is added. As shown in the right part of Figure 1, the graph has a positive effect on the performance of the models under different values of , which also increases with the proportion of outliers. In a situation where there are 40 % outliers in the training data, the graph even helps the test of LapPLR increase by over 0.6. This once again validates the claim of this study that through the introduction of graph Laplacian, data-driven models can learn in the correct direction when faced with polluted training data.

* + 1. Real-World High-Low Transformer Unit

The high-low transformer unit is a critical device in the real-world ammonia synthesis process, and its schematic diagram is shown in Figure 2. Within this unit, 26 process variables can be gathered, and the KQI variable is the residual content of CO at the outlet. To build the prediction model, 5000 training samples and 1500 testing samples were collected. The performance of different models is shown in Table 2.

Table 2. Best performance of models for the high-low transformer unit. The second column indicates the value of the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model |  | Training | Test | Test MSE |
| LR | N/A | 0.8960 | 0.7889 | 1.39 10-5 |
| MLP | N/A | 0.9547 | 0.4424 | 3.68 10-5 |
| LapSVR | 1.0 | 0.8408 | 0.7970 | 1.34 10-5 |
| LapHBR | 0.0 | 0.8954 | 0.7784 | 1.46 10-5 |
| **LapPLR** | **10.0** | **0.8589** | **0.8073** | **1.27 10-5** |

Both LapHBR and LapPLR exhibited better generalizability, with LapPLR performing the best. In addition, the relationship between the test and the graph Laplacian regularizer coefficient for LapHBR and LapPLR is also shown in Figure 3. Within the appropriate range, the test of the two models generally increases with .

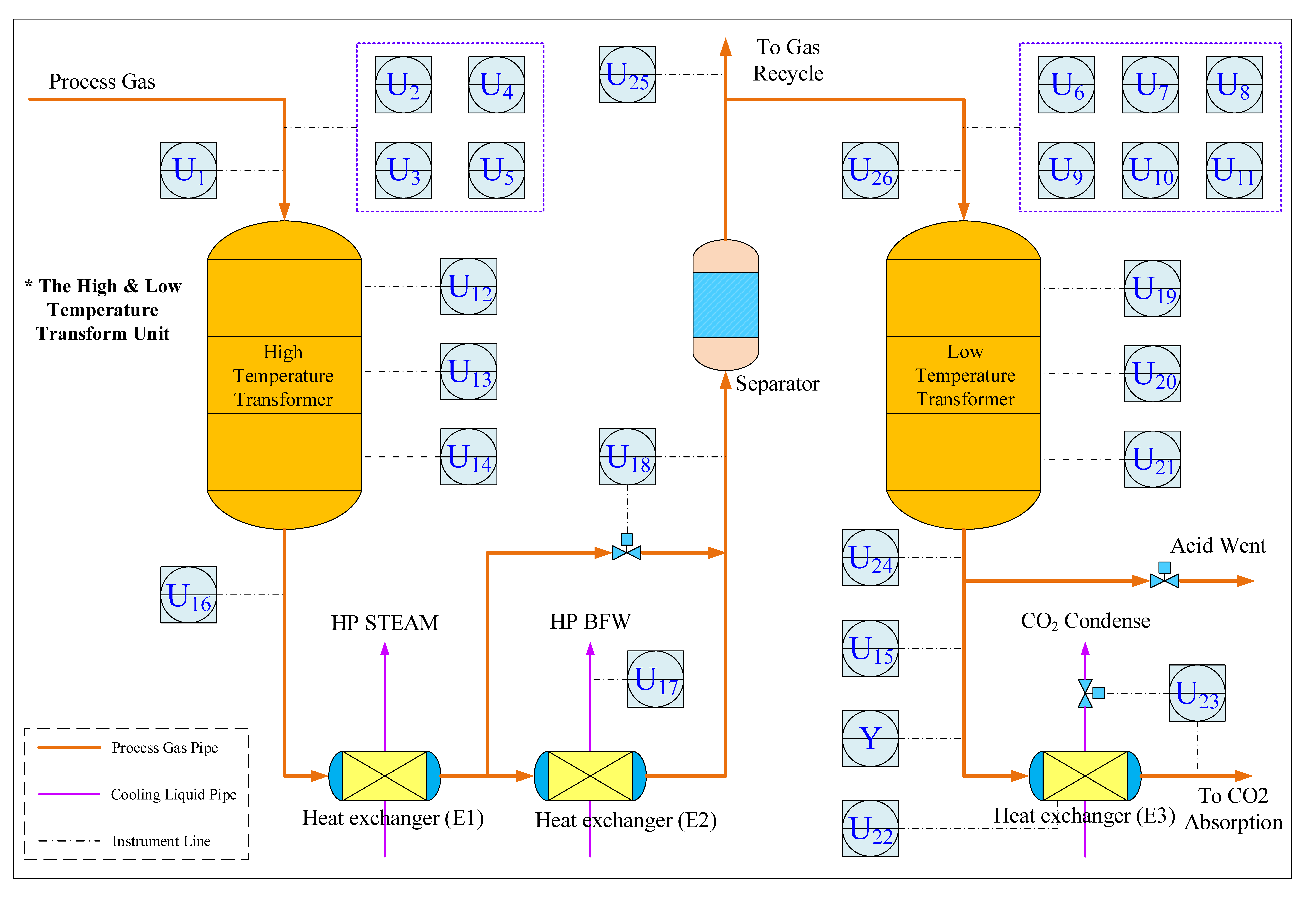
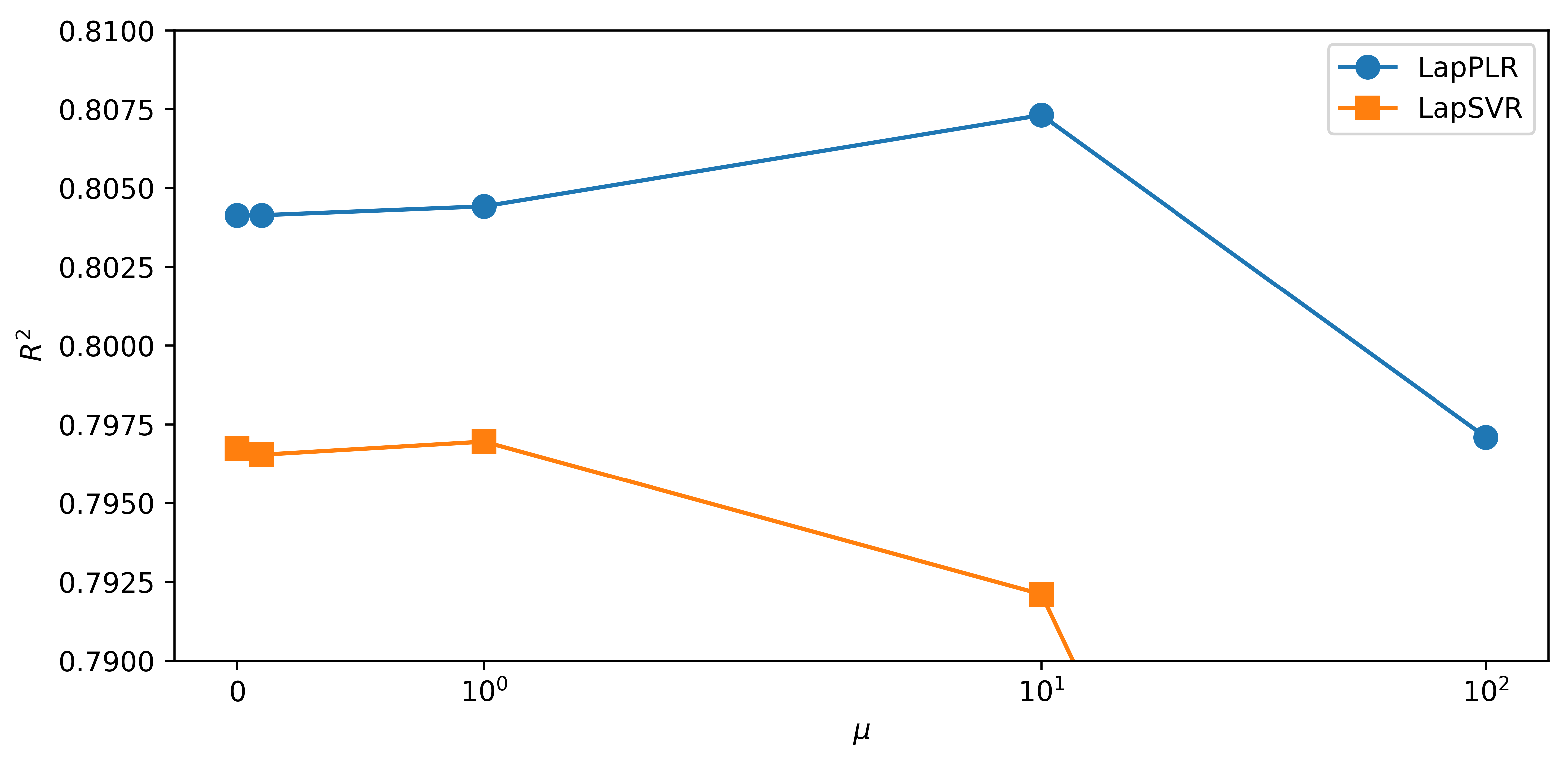
* 1. Conclusions

Figure 3. Test values of LapHBR and LapPLR with the change of model hyper-parameter .

Figure 2. The schematic diagram of the high-low transformer process.

To improve the robustness of the soft sensor model, two new soft sensor models LapHBR and LapPLR are proposed in this study, where Huber loss and piecewise linear loss are designed in the learning objective under the manifold regularization framework. In addition, dual problems for two new model objectives have been derived, which can be solved more efficiently by the quadratic programming solvers utilizing the interior-point methods compared to the primal model objectives. Besides, a general approach to obtain sample graph structure and the graph Laplacian used in the intrinsic regularizer has been proposed in this work. Such graph information can aid data-driven models, guiding them in the correct direction and at the same time keeping them from being misled by the noise or outliers inside the training data. Experiments have been conducted in both a simulation study and a real-world high-low transformer unit process, which demonstrated the excellent reliability of the proposed soft sensing model as well as the graph Laplacian constructed by the sample graph construction method.

References

Belkin, M., Niyogi, P. and Sindhwani, V., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of machine learning research, 7(11), 2399-2434.

Boyd, S. and Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, Cambridge, England.

Kano, M. and Fujiwara, K., 2013. Virtual sensing technology in process industries: trends and challenges revealed by recent industrial applications. Journal of chemical engineering of Japan, 46(1), 1-17.

Luo, Y., Zhang, X., Kano, M., Deng, L., Yang, C. and Song, Z., 2023. Data-driven soft sensors in blast furnace ironmaking: a survey. Frontiers of Information Technology & Electronic Engineering, 24(3), 327-354.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G. and Ljung, L., 2022. Regularized system identification: Learning dynamic models from data. Springer Nature, Cham, Switzerland.

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

Westerhuis, J.A., Derks, E.P., Hoefsloot, H.C. and Smilde, A.K., 2007. Grey component analysis. Journal of Chemometrics: A Journal of the Chemometrics Society, 21(10-11), 474-485.

Yao, L. and Ge, Z., 2017. Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. IEEE Transactions on Industrial Electronics, 65(2), 1490-1498.

Zhang, X., Kano, M. and Tani, M., 2023. Stacked supervised Poisson autoencoders-based soft-sensor for defects prediction in steelmaking process. Computers & Chemical Engineering, 172, 108182.