Data-driven dynamic inferential sensors based on causality analysis

Liang Cao[1], Feng Yu[2], Fan Yang[2], Yankai Cao[1], Bhushan Gopaluni[1*]

1. Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, Canada V6T 1Z3

2. Department of Automation, Tsinghua University, Beijing 100084, China

Abstract: Considering the stringent requirements for product quality of complex industrial processes, the purpose of this study is to apply causality analysis to select causal features of quality-relevant variables; and then to improve the prediction performance and interpretability of inferential sensors. Based on the idea that low-dimensional causal features can approximate the underlying information of the process instead of the original high-dimensional measurements, feature causality analysis is proposed in this work. To describe dynamic information and extract efficient latent features, dynamic latent variable models are utilized to combine with feature causality analysis. After dynamic latent causal feature extraction, two kinds of inferential sensors are developed with extracted dynamic latent causal features. Several comparison studies have been implemented on the Tennessee Eastman benchmark process; the results show that the inferential sensors based on dynamic latent causal features obtain the best performance.

Keywords: Inferential sensor; causality analysis; dynamic modeling; latent variable model.

1. Introduction

The process industry is an indispensable aspect of the global economy. With the increasing demands for higher product quality and cost efficiency, the complexity and automation of industrial processes is continuously growing [1]. As the complexity of industrial plants grows, the industrial plants face high risk of accidents [2]. Therefore, it is important to enhance the safety and reliability of process industry through process modeling and monitoring [1-8].

1.1 Background and Motivation

The inferential sensor (soft sensor) [4-7] provides a framework for dealing with imperfections in the measurements of complex processes using data-driven algorithms. In industrial processes, inferential sensor models are commonly developed to monitor variables that are associated with the quality of products, by establishing mathematical models between dependent variables (quality-relevant) and independent variables.

Feature selection (throughout the paper, for simplicity, features and variables are equivalent) is

an important first step in building inferential sensors, it is very complicated and often there is no unique of doing so. Causality analysis is a powerful tool to define whether a feature is 'important'. However, there are few applications of this technique in feature selection. In addition, traditional data-driven inferential sensors commonly rely on the assumption that processes operate at steady states and the data are temporally independent. However, there are often dynamic characteristics that define the relationship between the independent and dependent variables. Therefore, traditional steady-state inferential sensors are prone to inaccuracies and lack of robustness. When building inferential sensor model, introducing latent variable methods has some important benefits. One obvious benefit is to include features of interest that cannot be directly measured into process industry. These latent features are usually more efficient than the measurable variables in developing inferential sensors.

In this work, dynamic latent variable models are utilized to extract dynamic latent features. In addition, to select efficient latent features from the extracted latent features, we proposed two causality analysis methods to describe causality relationship between the extracted latent features and quality-relevant variables and select latent features that have high causality with quality-relevant variables. Finally, two kinds of inferential sensors are developed to address the above-mentioned problem with dynamic latent causal features.

1.2 Literature Review

This section contains a detailed literature survey on recent work for dynamic modeling, feature selection, causality analysis and latent variable model.

To effectively model dynamic characteristics that exist in industrial process, several methods have been proposed and they can be classified into three categories: vector augmentation by lagged samples [5], recurrent neural network [7] and system identification [6]. Vector augmentation by lagged samples is one common approach to model dynamics in inferential sensors. System identification is another approach that has been successfully used in process control and monitoring. In [6], an impulse response template model based on Wiener structure is used to describe the dynamic relationship between variables by introducing impulse response function. Recurrent neural network is an example of deep learning, and it has also been successfully used in dynamic inferential sensors.

Currently, most common feature selection approach is search and score, it defines score function $f_s : R^n \mapsto R$ that measures quality of a set of features $S = \{x_j\}_{j=1}^{s}$ ($X \in R^{n \times m}$ and $y \in R^n$ are the

independent variables and dependent variables , s is the number of selected features) and then search for the set S with best score; L_0 regularization and L_1 regularization are the most common score functions [9]; forward selection and backward selection are the most common search methods [10]. However, this approach is usually computationally expensive on real process data due to the large number of sets of variables. It is also difficult to define an appropriate 'score' function and 'search'.

In inferential sensors, we usually identify causal effects between independent variables and dependent variables by "forcing" the independent variable to take a certain value and then measuring the effect in dependent variable. If there is strong dependence between the variables, it indicates that the corresponding independent variable is important for predicting the dependent variable and there is a causal effect. With the ability to mine the causal dependency, causality analysis can provide valuable information about the process when expert knowledge is limited. Causality analysis [11-18] has played a key role in alarm root causality analysis and fault diagnosis [13-15]. However, there are few applications of this technique in inferential sensors. To develop a more interpretable and more accurate inferential sensor while maintaining the simplicity of the model, this work uses causality analysis to capture causality hidden in process data; and exploit features that have causal relationship with dependent variables to develop inferential sensors.

For time series process data, there are several effective methods (time delay, conditional probabilities, energy transfer) to capture and quantify causality. These methods can be classified into parametric and nonparametric, linear and nonlinear, etc. For simplicity, we use one of the methods proposed in [18] to classify causality analysis methods into two categories: linear and nonlinear.

Among these methods, Granger causality analysis (GCA) [11-12] and transfer entropy (TE) [13-14,16-17] are two of the most common methods for establishing pairwise causality of variables and therefore have been used in many areas, like economics, biology and process industries. In [11], the author concluded that TE theoretically provides more information about the causal relationship especially when the data are limited, but both methods generally exhibit similar performance with enough data. It needs to be pointed that many other effective methods [19-20] are also reported with the increasing attention on causality analysis, which are not covered in this work.

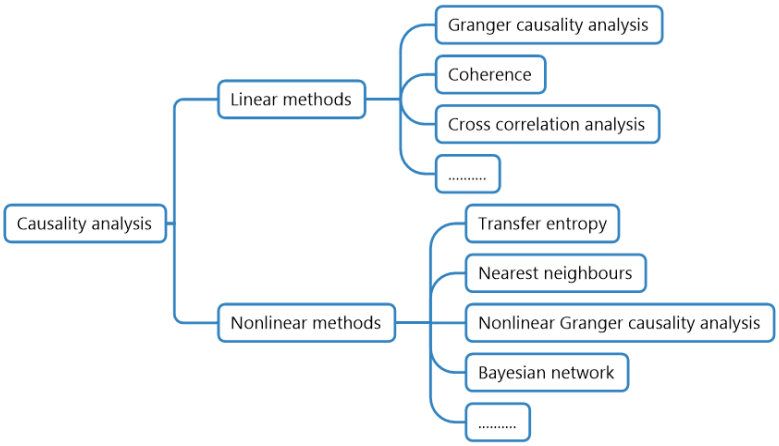

Figure1. Classification of causality analysis methods

A ‘latent variable’ in multivariate statistical process monitoring is a random variable that is unmeasured but contains abundant information about process data. Many multivariate statistical models use latent variables such as principal component analysis (PCA) [8], slow feature analysis (SFA) [1], canonical correlation analysis (CCA) [8], partial least square (PLS) [5]. In addition, some nonlinear latent variable models, like variational autoencoder (VAE) [21] and stacked autoencoder (SAE) [22] have also been proposed and applied in inferential sensors.

1.3 Contribution and Paper Organization

The present work aims to develop a more interpretable and more accurate inferential sensor while maintaining the simplicity of the model. For a high-dimensional industrial process, it is important to find a subset of the ‘important’ features to predict dependent variables as more variables imply the complex model while less variables mean the loss of useful information. As an important contribution, this work uses causality analysis to capture causality hidden in process data; and exploit features that have high causality relationship with dependent variables to develop inferential sensors. Based on this idea, two types of advanced dynamic latent causal inferential sensors are proposed. Unlike traditional methods which either cannot provide precise prediction using simple model or suffer from high model complexity; the model complexity of this proposed algorithm is reduced significantly by introducing causality analysis and latent variable model. The case study shows that proposed inferential sensors are highly efficient in improving prediction performance and reduce model complexity.

This article is organized as follows. Section 2 reviews algorithms of causality analysis, dynamic

inferential sensor, as well as latent variable model and regularization methods. In Section 3, new approaches for proposed dynamic latent causal inferential sensors are put forward, with detailed implementation procedures and algorithmic analysis. Section 4 gives comparison case studies on the Tennessee Eastman benchmark process to verify the effectiveness of proposed methodology. Concluding remarks are presented in section 5.

2. Algorithm Principle

Our intended problem is to find a subset of dynamic causal latent features, so that an accurate and simple inferential sensor can be easily found. Several algorithms and methods about causality analysis, dynamic modeling and latent variable model will be given comprehensive discussion in this section.

2.1. Causality Analysis

Based on causality analysis categories that mentioned in introduction**,** two causality analysis methods, Granger causality analysis (linear) and transfer entropy (nonlinear) will be introduced in the following section.

A. Granger Causality Analysis

The idea of Granger causality originally came from Wiener, which could be expressed as: time series $x(t) \in R^n$ causes $y(t) \in R^n$ if the predictability of $y(t)$ could be improved by introducing the information of $x(t)$. Granger [12] used a bivariate auto-regressive model to formalize the idea, that is, the introduction of the historical information of $x(t)$ reduces the prediction error of $y(t)$ in a bivariate autoregressive model compared with the model only using the past information of $y(t)$. Supposing that $x(t)$ and $y(t)$ could be modeled as an autoregressive model:

$$\begin{aligned} x(t) &= \sum_{d=1}^{p} a_{1,d} x(t-d) + \varepsilon_{1x}(t), \operatorname{var}\left(\varepsilon_{1x}(t)\right) = \Gamma_{1x} \\ y(t) &= \sum_{d=1}^{p} b_{1,d} y(t-d) + \varepsilon_{1y}(t), \operatorname{var}\left(\varepsilon_{1y}(t)\right) = \Gamma_{1y} \end{aligned} \tag{1}$$

Jointly, they could be described by a bivariate autoregressive model:

$$\begin{aligned} x(t) &= \sum_{d=1}^{p} a_{21,d} x(t-d) + \sum_{d=1}^{p} a_{22,d} y(t-d) + \varepsilon_{2x}(t), \operatorname{var}\left(\varepsilon_{2x}(t)\right) = \Gamma_{2x} \\ y(t) &= \sum_{d=1}^{p} b_{21,d} x(t-d) + \sum_{d=1}^{p} b_{22,d} y(t-d) + \varepsilon_{2y}(t), \operatorname{var}\left(\varepsilon_{2y}(t)\right) = \Gamma_{2y} \end{aligned} \tag{2}$$

where scalar p is the model order, Γ_{1y} is the variance of residual $\varepsilon_{1y}(t)$. Γ_{1y} measures prediction accuracy of $y(t)$ using its own information while Γ_{2y} measures prediction accuracy of $y(t)$ using information of both $x(t)$ and $y(t)$. According to the definition of Granger causality, if Γ_{2y} is less than Γ_{1y}, then it is concluded that $x(t)$ causes $y(t)$. The strength of causal connectivity can be measured as follows:

$$G_{x(t)\to y(t)} = \ln \frac{\Gamma_{1y}}{\Gamma_{2y}} \tag{3}$$

The figure 2 gives an intuitive explanation of Granger causality. In Granger causality analysis, statistical significance tests need to be performed before a causal link can be established. After the statistical significance test, a causal map could be built between each variable.

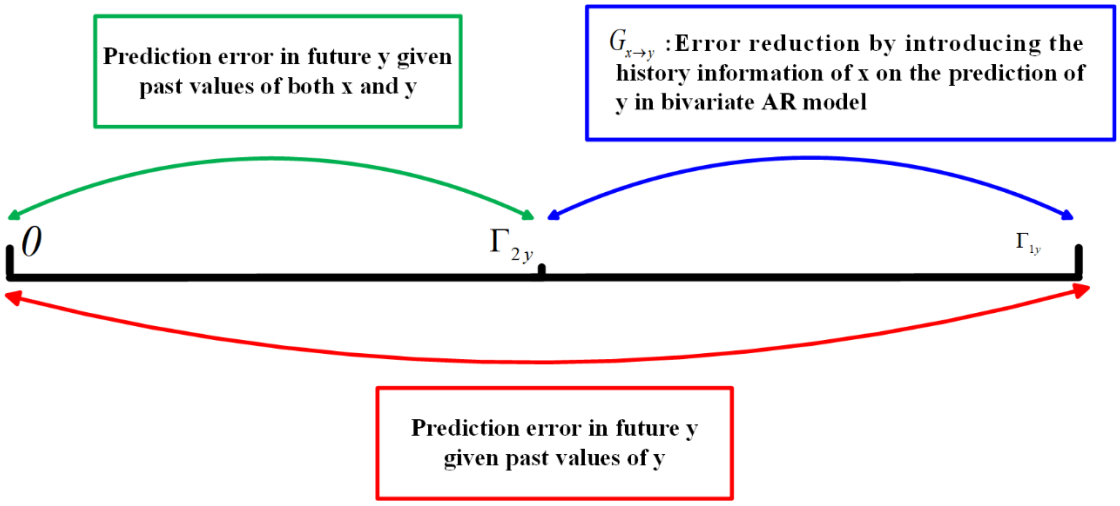

Figure 2. Intuitive explanation of Granger causality

B. Transfer Entropy

It is well known that entropy measures "randomness" of a set of variables, low entropy means "very predictable" and high entropy means "very random". Based on this idea, transfer entropy was proposed by Schreiber to measure the uncertainty reduction between variables [16]. Namely, transfer entropy measures the uncertainty reduction in future y given past information of x and y, compared with only using the past information of y.

Supposing that $x(t)=\left[x(1),\cdots,x(i),\cdots,x(n)\right]^T \in R^n$, $y(t)=\left[y(1),\cdots,y(i),\cdots,y(n)\right]^T \in R^n$, and the transfer entropy from $x(t)$ to $y(t)$ can be calculated as follows:

$$T_{x(t)\to y(t)} = \sum_{y_{t+h},y_t^{(k)},x_t^{(l)}} p\left(y_{t+h},y_t^{(k)},x_t^{(l)}\right)\log\frac{p\left(y_{t+h}\left|y_t^{(k)},x_t^{(l)}\right.\right)}{p\left(y_{t+h}\left|y_t^{(k)}\right.\right)}$$

$$= \sum_{y_{t+h},y_t^{(k)},x_t^{(l)}} p\left(y_{t+h},y_t^{(k)},x_t^{(l)}\right)\log p\left(y_{t+h}\left|y_t^{(k)},x_t^{(l)}\right.\right) - \sum_{y_{t+h},y_t^{(k)},x_t^{(l)}} p\left(y_{t+h},y_t^{(k)},x_t^{(l)}\right)\log p\left(y_{t+h}\left|y_t^{(k)}\right.\right) \quad (4)$$

$$\triangleq H\left(y_{t+h}\left|y_t^{(k)}\right.\right) - H\left(y_{t+h}\left|y_t^{(k)},x_t^{(l)}\right.\right)$$

where $p(\cdot)$ is the complete or conditional probability density function (PDF), l,k are the order of variables $x(t),y(t),h$ is the prediction horizon. $x_t^{(j)}=\left[x(t-j),\cdots,x(t-1),x(t)\right]^T$ and similar for $y_t^{(j)}$.

The figure 3 gives an intuitive explanation of transfer entropy. Based on the definition of transfer entropy in (4), we can conclude that, if there is uncertainty reduction in $y(t)$ given past values of $x(t)$, we would say that $x(t)$ causes $y(t)$ and the value of $T_{x(t)\to y(t)} > 0$. As with Grange causality analysis (GCA), a significance test is also needed to determine whether there is a causal relationship between variables. The significance threshold could be obtained by using a Monte Carlo method with surrogate data [17].

The TE method is a type of nonlinear causality analysis approach based on information theory and has a wide range of applications in causality analysis. It is based on probability theory and does not depend on the model, so we can directly use this method to find important independent variables that have pairwise causal relationships with dependent variables.

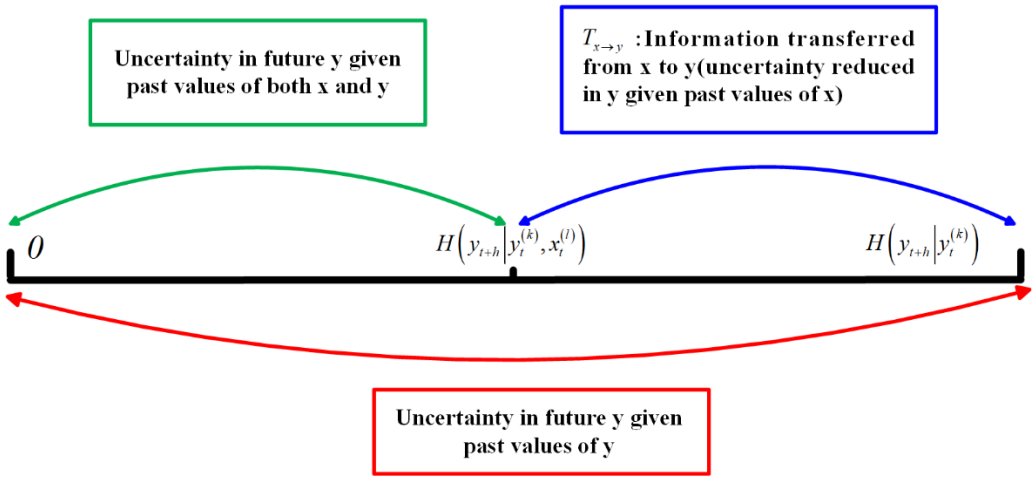

Figure 3. Intuitive explanation of transfer entropy

2.2 Dynamic Inferential Sensor

For independent variables $X=\left[x_1(t),x_2(t),...,x_m(t)\right]\in R^{n\times m}$ and dependent variable $y(t)=\left[y(1),\cdots,y(i),\cdots,y(n)\right]^T\in R^n$, assume that $x_k(t)=\left[x_k(1),\cdots,x_k(i),\cdots,x_k(n)\right]^T\in R^n$ is the kth variable, where $x_k(i)$ is the value of kth variable at time i. When using vector augmentation by time lag d on $x_k(i)$, one gets:

$$\mathbf{x}_k(i)=\left[x_k(i),\quad x_k(i-\Delta t),\quad \cdots,\quad x_k(i-d\Delta t)\right]\in R^{d+1},1\le i\le n \tag{5}$$

Before vector augmentation by lagged samples, the vector of independent variable at time i is $X_i=\left[x_1(i),x_2(i),...,x_m(i)\right]\in R^m$, $1\le i\le n$. After vector augmentation by lagged samples, as figure 4 shows, the vector becomes $X_i=\left[\mathbf{x}_1(i),\mathbf{x}_2(i),\cdots,\mathbf{x}_m(i)\right]\in R^{m(d+1)}$. Therefore, for historical observations with time lag d, the input matrix is changed as $X\in R^{n\times m(d+1)}$. Here, if we use $X\in R^{n\times m(d+1)}$ as input and choose least square regression to predict $y(t)$, then we can get the algorithm called dynamic least square regression and this algorithm will be used in case study section.

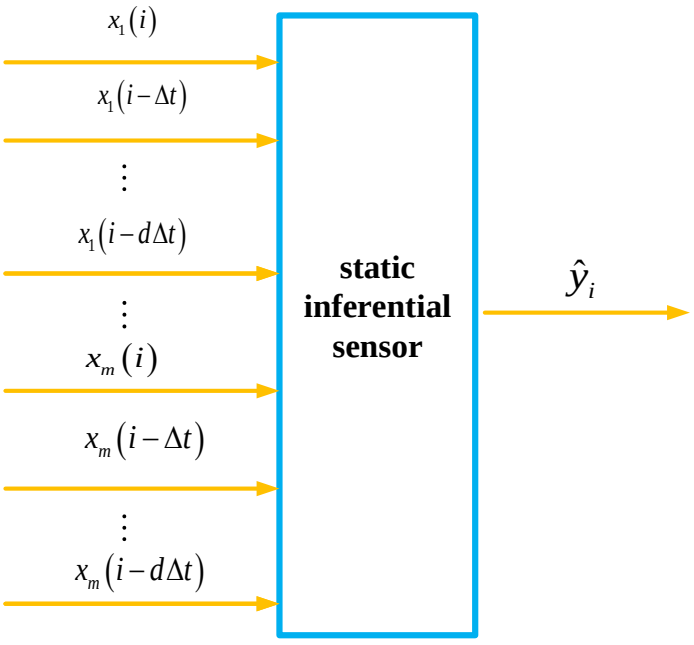

Figure 4. The structure of vector augment by lagged samples method

2.3 Latent Variable Analysis

In the following section, we give brief introduction on PCA and SFA that used in this work.

A. Principal Component Analysis

Principal component analysis (PCA) is a classical feature extraction method to find those latent variables that carry the most variance information from the original data. For a standardized data matrix $X \in R^{n\times m}$, PCA tries to decompose X as the following equations:

$$X = TP^T + E \tag{6}$$

where $T = [t_1, t_2, ..., t_\lambda] \in R^{n\times\lambda}$ is the score matrix, $P \in R^{m\times\lambda}$ is the loading matrix, $t_i \in R^n$ is the latent variable known as the principal component, $E \in R^{n\times m}$ is the residual matrix. The computation of PCA can be done with singular value decomposition, gradient descent, etc. λ is the number of principal components and it could be selected according to the methods proposed in [23].

B. Slow Feature Analysis

Industrial processes have significant inertial characteristics and usually operate steadily, so they possess slow dynamic characteristics in most cases. The central idea of slow feature analysis is to extract the slowest components from the time series data and treat them as main features. The slow changes can be assumed to represent the fundamental characteristic features of a process, while the fast changes can be treated as short-term noise.

For a given time series signal $x(t) \in R^n$, the speed of change can be measured by $\Delta(x(t)) = \langle \dot{x}^2(t) \rangle_t$, $\dot{x}(t) = x(t) - x(t-1)$ is the time difference and $\langle x(t) \rangle_t = \frac{1}{n}\sum_{i=1}^{n} x(t_i)$ denotes the time averaging value of a certain time series with n observations. For data matrix $X(t) = [x_1(t), x_2(t), ..., x_m(t)]$, the purpose of SFA is to find a set of slow features $S(t) = [s_1(t), s_2(t), ..., s_m(t)]$ such that the slowness of extracted feature is minimal. The objective of this optimal problem is:

$$\begin{aligned} &\min_{g_i(\cdot)} \langle \dot{s}_i^2 \rangle_t, i = 1, ..., m \\ &s.t. \ \langle s_i \rangle_t = 0 (\text{zero mean}), \langle s^2{}_i \rangle_t = 1 (\text{unit variance}) \\ &\forall i \neq j, \langle s_i s_j \rangle_t = 0 (\text{decorrelation and order}) \end{aligned} \tag{7}$$

where $g_i(\cdot)$ is the scalar function that needs to be optimized, $s_i(t) = w_i^T X(t)$ for linear function. Shang [1] gave a detailed introduction of the solutions of SFA problem. To determine the number of slow features, a q-upper quantile value of the slowness of inputs is introduced as follows:

$$M_e = card\left\{ s_i \mid \Delta s_i > \max_j^q \left\{ \Delta x_j \right\} \right\} \tag{8}$$

where $\text{card}\{\cdot\}$ represents the number of elements in set and M_e is the number of features that are ought to be removed.

2.4. Elastic Net

In this work, we will use elastic net as one of the comparison studies to show the effectiveness of causality analysis in feature selection of inferential sensors. Elastic net [24] is the combination of L_2 regularization (ridge regression) and L_1 regularization (Lasso). L_2 regularization adds penalty on the L_2 norm of and L_1 regularization adds penalty on the L_1 norm of w to generate sparse solutions and select features. For nonnegative λ and $\sigma \in (0,1)$, the cost function of elastic net based on least square can be written as:

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \lambda\left(\frac{1-\sigma}{2}\|w\|^2 + \sigma\|w\|_1 \right) \tag{9}$$

This cost function is the same as the cost function of ridge regression when σ=0 and Lasso when σ=1. For L_1 regularization, it is robust when there are a number of irrelevant features as those features are ignored (many w_j tend to be zero) in the prediction of y. But the solution of L_1 regularization is not unique as there may be multiple w values that achieve the minimum. L_2 regularization can provide unique solution as it is strongly convex. Therefore, by combing L_2 and L_1 regularization, the elastic net can simultaneously select features and provide sparse, unique solutions.

Remark: It is worth pointing out that under the assumption of Gaussian likelihood and Gaussian prior, the maximum a posteriori (MAP) estimation is equivalent to solving the L_2 regularized least squares problem in the "loss plus regularization" framework. Similarly, Laplace prior leads to L_1 regularization. The choice of prior corresponds to the choice of regularization and this requires us to give plausible assumption about w in real applications when using regularization.

3. Proposed Methodology

As mentioned in introduction, it is extremely necessary to find a subset of the 'important' measurable features or latent features that are well-defined and interpretable to dependent variables. The high-dimensional measurements in industrial processes are often highly correlated

[8]. There is often a low-dimensional feature space that explains the most important information in observations [25]. In other words, some causal features could fully represent the underlying information of the process instead of all the original variables.

In this section, a dynamic inferential sensor based on causality analysis is proposed using the idea of feature learning in industrial processes. To develop a more interpretable and more accurate inferential sensor, this work uses causality analysis to capture causality hidden in process data; and exploit features that have causal relationships with dependent variables to develop inferential sensors. After feature causality analysis, the causal features could be selected as the independent variables to predict dependent variables. A detailed description of proposed methodology is given in the following section.

3.1. **Causal Feature Learning of Dynamic Inferential Sensor**

For a high-dimensional industrial process, it's hard to select appropriate independent variables to predict dependent variables due to the trade off between the complex of model and the loss of useful information. Obviously, a set of variables which have causal influence on dependent variables are the most important variables. Using this idea, causal feature learning of dynamic inferential sensor is proposed in this work.

Figure 5 shows the framework of causal feature learning of dynamic inferential sensor. Vector augmentation by lagged samples is performed initially to obtain dynamic features $X^d \in R^{n\times m(d+1)}$ of original process data. Then, given dependent variable to be analyzed, causality analysis is conducted to select the causal features $\tilde{X}^d \in R^{n\times c}$ $\left(\tilde{x}_i^d(t), 1\le i\le c < m(d+1)\right)$ by measuring their causal influence on the dependent variables because an important feature should exert more causal influence on the dependent variables. The causal features $\tilde{X}^d$ are the subset of features X^d. Then, latent variable model is performed to extract latent features $Z \in R^{n\times l}$ $\left(z_i(t), 1\le i\le l\right)$ of causal features. With latent causal features, to build dynamic model for inferential sensor, dynamic latent causal features $Z^{d^*} \in R^{n\times l\left(d^*+1\right)}$ $\left(z_i^{d^*}(t), 1\le i\le l\left(d^*+1\right)\right)$ are obtained as inputs to the inferential sensor.

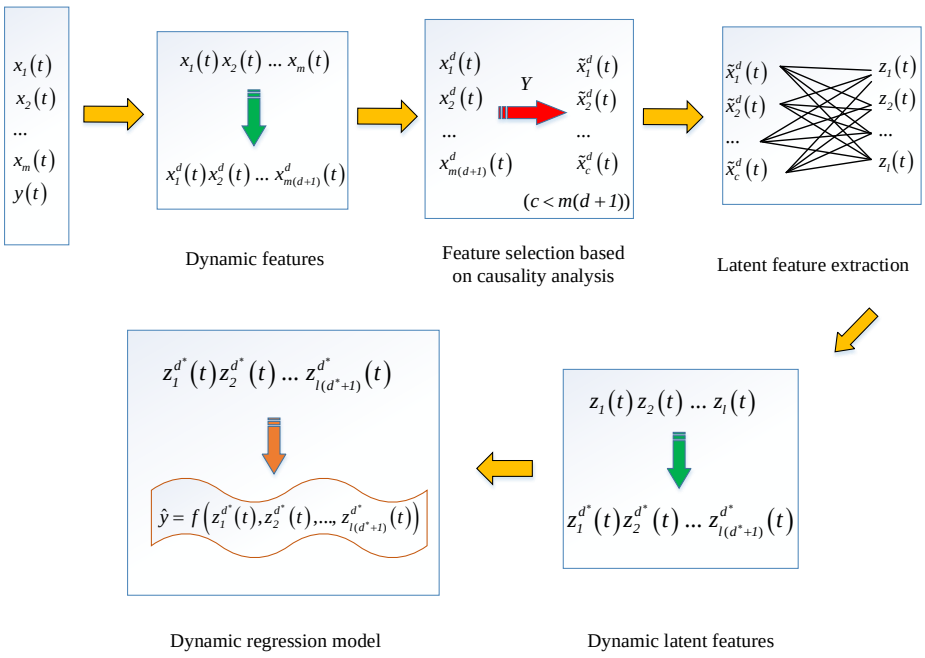

Figure 5. causal feature learning framework

3.2. **Latent causal feature Learning of Dynamic Inferential Sensor**

In some cases, process data possesses strong correlation and collinearity. Hence, finding directed causality between original independent variables and the dependent variables may be hard and complicated. However, there is no correlation between each feature in most latent variable models like SFA, PCA; as a result, small causal links could be assumed among these latent features, which is useful for selecting latent features in building inferential sensors.

Figure 6 illustrates how latent causal feature learning of dynamic inferential sensors is performed. Like the approach in section 3.1, vector augmentation by lagged samples is initially performed to obtain dynamic features X^d of the original data. Then, latent variable model is used to extract essential latent features $\tilde{Z} \in R^{n\times l}\left(\tilde{z}_i(t), 1 \le i \le l\right)$ of dynamic features. Then, given dependent variable to be analyzed, causality analysis is conducted to select the latent causal features $Z \in R^{n\times c}\left(z_i(t), 1 \le i \le c \le l\right)$ by measuring pairwise causality between latent features and dependent variables. Similarly, the latent causal features Z are the subset of latent features $\tilde{Z}$. After determining all the important latent causal features, to build dynamic model for inferential sensor, dynamic latent causal features $Z^{d^*} \in R^{n\times c\left(d^*+1\right)}\left(z_i^{d^*}(t), 1 \le i \le c\left(d^*+1\right)\right)$ are obtained as inputs of inferential sensor.

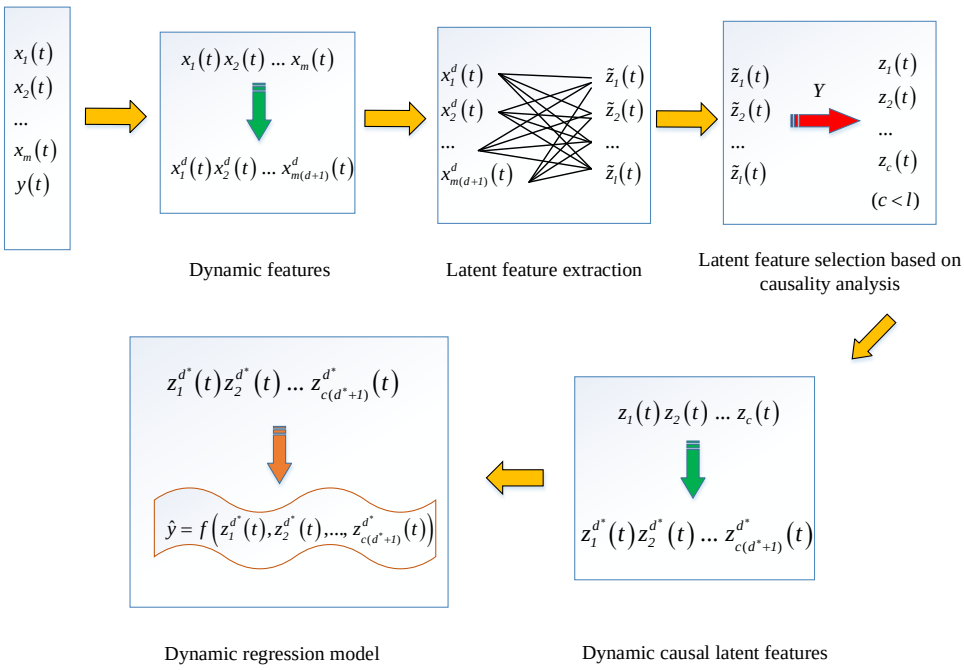

Figure 6. latent causal feature learning framework

3.3. **Inferential Sensor Design and Analysis**

For the proposed algorithms, the major difference between the two proposed dynamic inferential sensors is the order of causality analysis and latent feature learning. The implementation of these two algorithms can be split into four stages.

1. Dynamic feature extraction: In this step, data preprocessing will be performed to deal with the missing values, data outliers and drifting data. After data preprocessing, choose appropriate lagged time d for all independent variables and get the initial dynamic features for inferential sensors.

Remark: Considering the different structures of two proposed algorithms, we can find that the stage 2 and 3 of them are roughly opposite, which means the stage 2 of causal feature learning algorithm is the stage 3 of latent causal feature learning algorithm and vice versa. For simplicity, we only introduce stage 2 and 3 of the latent causal feature learning algorithms.

2. Latent feature extraction: Some efficient latent feature extraction algorithms, like PCA and SFA, could be used in this stage. As dynamic features of independent variables have been obtained in stage 1, we can apply dynamic PCA, dynamic SFA or other dynamic latent variable models to extract dynamic latent features.

3. Causality analysis: To select appropriate latent features for regression model, we could utilize causality analysis methods, such as GCA, TE or Bayesian network, to measure causality and select latent features that have strong causality with dependent variables.

4. Dynamic regression model: After selecting latent causal features, appropriate lagged time d^* for these latent causal features is chosen in stage 4 to get dynamic latent causal features. Some regression methods, such as least square, support vector regression or neural network, could be used to develop model.

Figure 7 shows the flowchart of two different structures of inferential sensor design based on causality analysis. For these two dynamic inferential sensors, red 1 represents causal feature learning while blue 2 represents latent causal feature learning. For offline learning, two feature learning methods are applied to get the dynamic latent causal features of dependent variables. Once a model has been developed and thoroughly tested offline with tons of data, it needs to be validated and evaluated online with real time data. In online monitoring, latent feature extracting algorithm and causality analysis methods learned in offline learning are applied to real time process data. Latent causal features could be extracted according to their indices in offline learning. With trained model parameters, assuming dynamic latent causal features as the inputs of the inferential sensor, real time predicted values of dependent variables could be obtained.

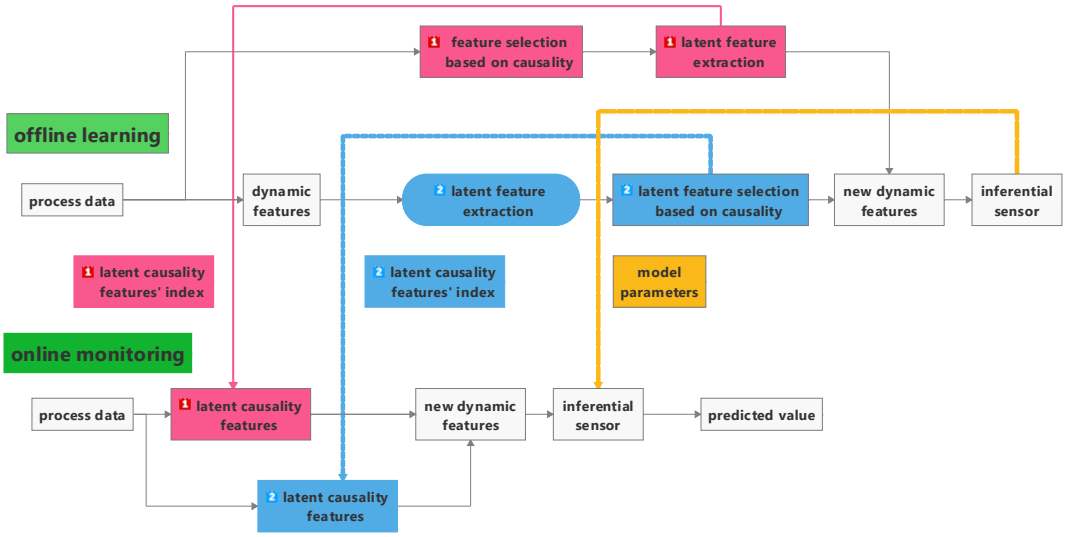

Fig 7. The flowchart of inferential sensor based on causality analysis

4. Case Study

Tennessee Eastman Problem (TEP) provides a benchmark for realistic industrial process monitoring. In this section, the effectiveness of proposed methodology is illustrated based on TEP [26]. Figure 8 shows the diagram of the TEP process. There are 52 different variables in this process, among which 33 variables can be measured in real time while another 19 variables need to be analyzed respectively.

In this case study, 33 variables are chosen as the independent variables and XMEAS (31), component C in purge gas, is chosen as the dependent variable to be predicted. The normal condition data is used to verify proposed methods, of which 960 samples for training and 500 samples for testing.

For dynamic independent variables, two causality analysis methods, Granger causality analysis (linear) and transfer entropy (nonlinear), are used to select features that have high causality relationships with dependent variables; two latent variable models, PCA and SFA, are used to extract latent features. Finally, dynamic least square regression method is used to build inferential sensors. In addition, elastic net and basic dynamic least square regression (DLSR) are performed for comparison studies. Correlation coefficient (r) between the predicted values and real values plus a root mean square error (RMSE) are calculated to show the performance.

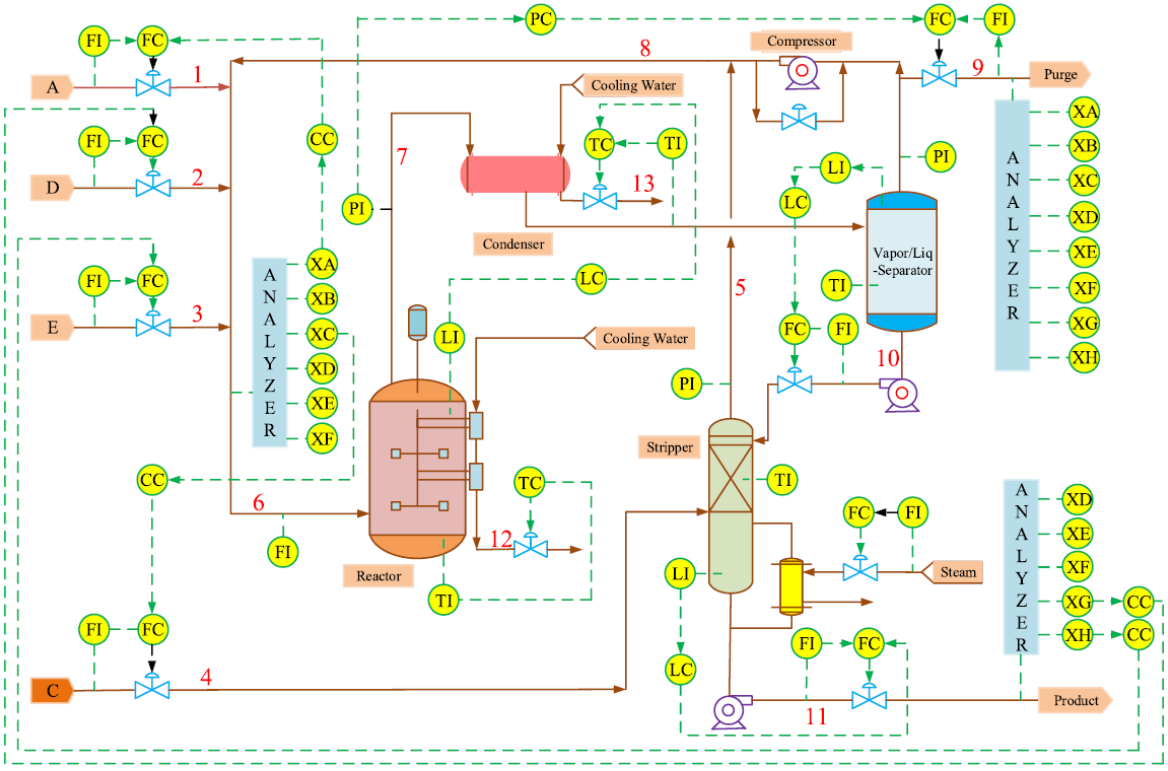

Figure 8. TEP flowchart

The detailed coefficients of each part of proposed algorithms are given as follows. The time lag d is set to 1 for original independent variables, so 33 independent variables will become 66 dynamic independent variables. The number of principal components is set when cumulative percent variance is greater than 0.85 and the number of slow features is determined by q-upper quantile value which is set to 0.1. For Granger causality analysis, history length is set to 5 and the

significance level is set to 0.05. The mean of TE value obtained by surrogate data is set as significance threshold of TE. The time lag d^* of regression model is set as 5; the coefficient λ of elastic net is selected by using 5-fold cross-validation and the σ is set as 0.75.

4.1. Results of Causality Analysis

A. Granger Causality Analysis

In this section, in order to give a complete demonstration of the effectiveness of our method, GCA is implemented to find the causality relationship between 66 dynamic features and all 19 dependent variables. We proposed two different algorithms, feature learning and latent feature learning, to extract dynamic latent causal features.

Firstly, the results of latent causal feature learning based on GCA are given in fig 9-11. Fig 9 shows that for all 19 dependent variables, the number of causal features is less than 20 and the number of latent causal features is less than 15, which means sharp decreases in the number of inferential sensor inputs.

Figure 10 and 11 show maximum and average causal connectivity strength of 19 variables respectively. High average and maximum causal connectivity strength are expected when evaluating the performance of proposed algorithms. It is obvious that for different dependent variables, the performance of PCA and SFA is different; PCA outperforms SFA in some cases while SFA outperforms PCA in other cases.

In the following part, we will focus on dependent variable 31. The green box in fig 9 shows different numbers of causal features and latent causal features of variable 31. To compare the input number of different inferential sensors, table 1 lists the number of inputs used in GCA latent causal feature learning algorithm and original inferential sensor without causality analysis. In addition, considering elastic net could select features and reduce the model complexity, the number of inputs used in elastic net is also included in table 1. Table 2 lists the number of inputs used in GCA causal feature learning algorithm.

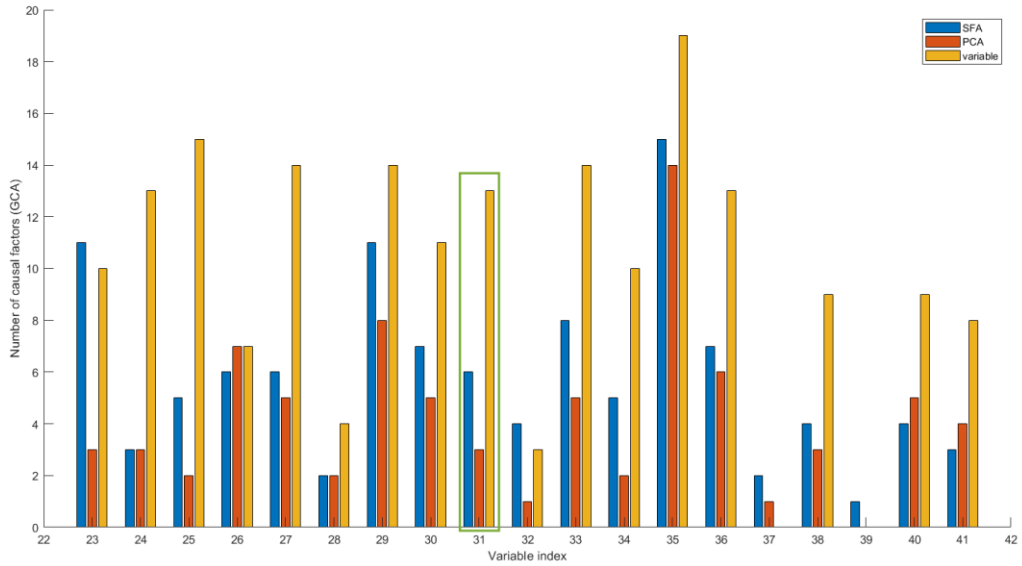

Figure 9. Number of causal features in GCA

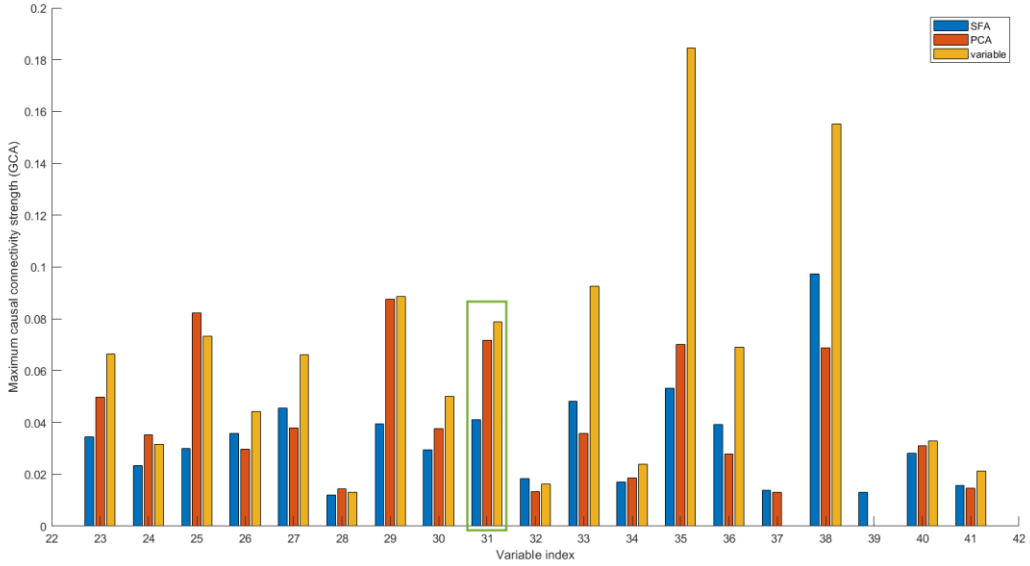

Figure 10. Maximum causal connectivity strength in GCA

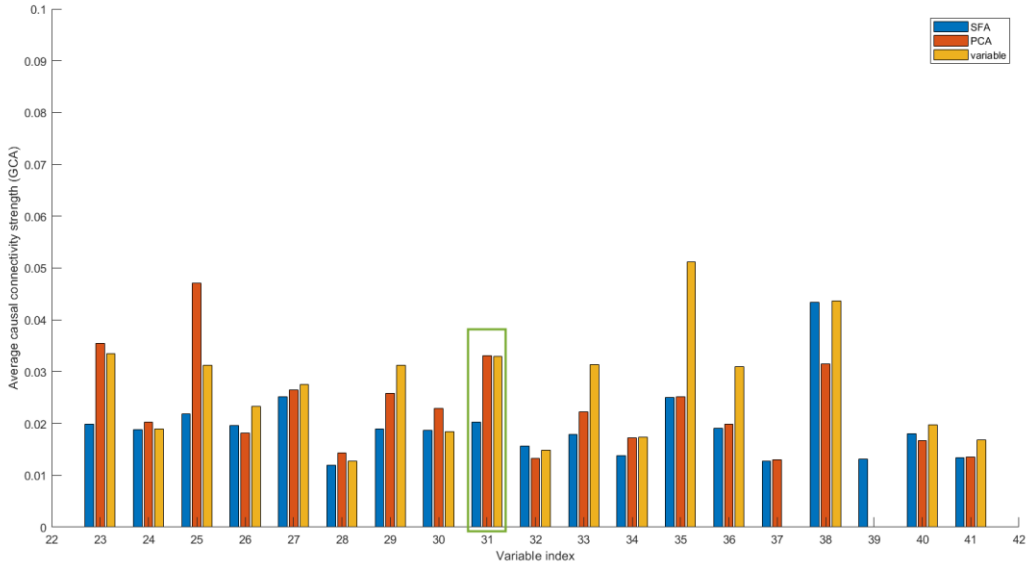

Figure 11. Average causal connectivity strength in GCA

Table 1. Number of inputs in GCA latent causal feature learning

Inferential sensor	Number of inputs
PCA+GCA	**3**
PCA	24
SFA+GCA	**6**
SFA	38
variable+GCA	**13**
variable	66
elastic net	17

Table 2. Number of inputs in GCA causal feature learning

Inferential sensor	Number of inputs
GCA+PCA	**9**
GCA+SFA	**7**
variable+GCA	**13**

B. Transfer Entropy

In this section, nonlinear causality analysis method, transfer entropy is implemented to find the causality relationship between 66 dynamic features and dependent variable. The causality relationship between them can be seen from figure 12. The white area represents that the value of the transfer entropy failed to pass the significance test. The red area represents that there is a causality relationship between dynamic feature and dependent variable, and the deeper color means the stronger causality.

Similar with GCA, two different schemes are applied to extract dynamic latent causal features. The results of latent causal feature learning based on TE are given in fig 12-14. Fig 12 shows the number of causal features of dependent variable is 28. Fig 13-14 show the numbers of PCA, SFA causal features are 1 and 2 respectively. Furthermore, we can find that the latent causal feature of PCA in inferential sensor is the first principal components and the latent causal features of SFA in inferential sensor are the first slowest and third slowest components.

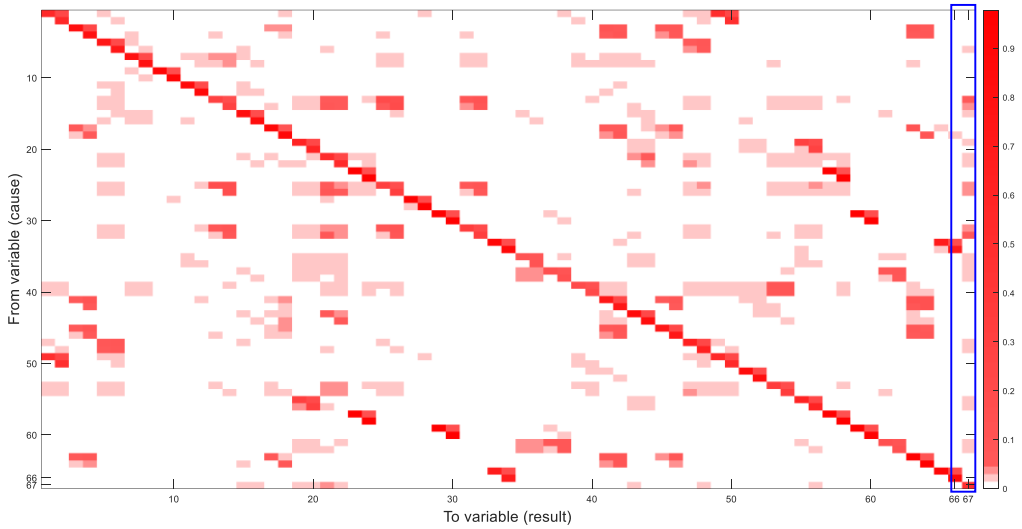

Figure 12. TE analysis for all dynamic features

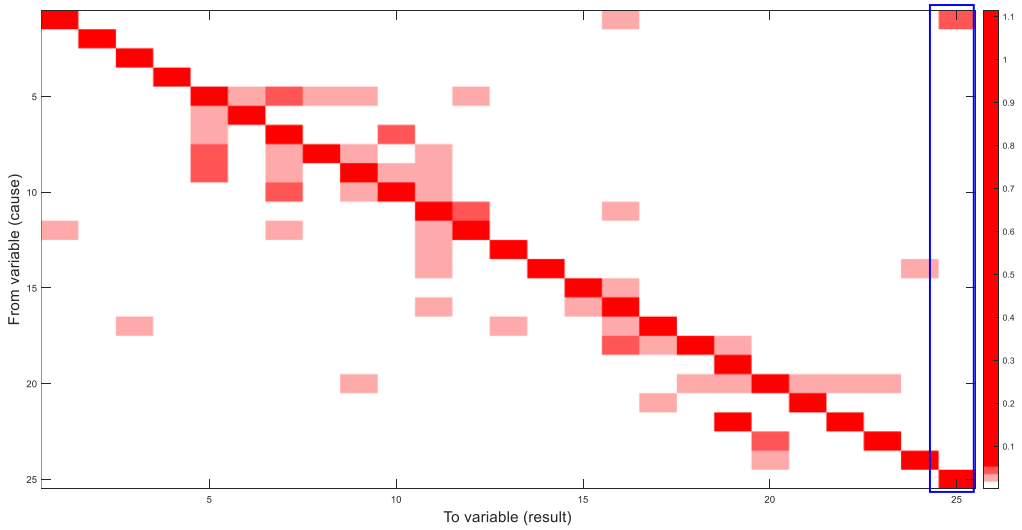

Figure 13. TE analysis for PCA

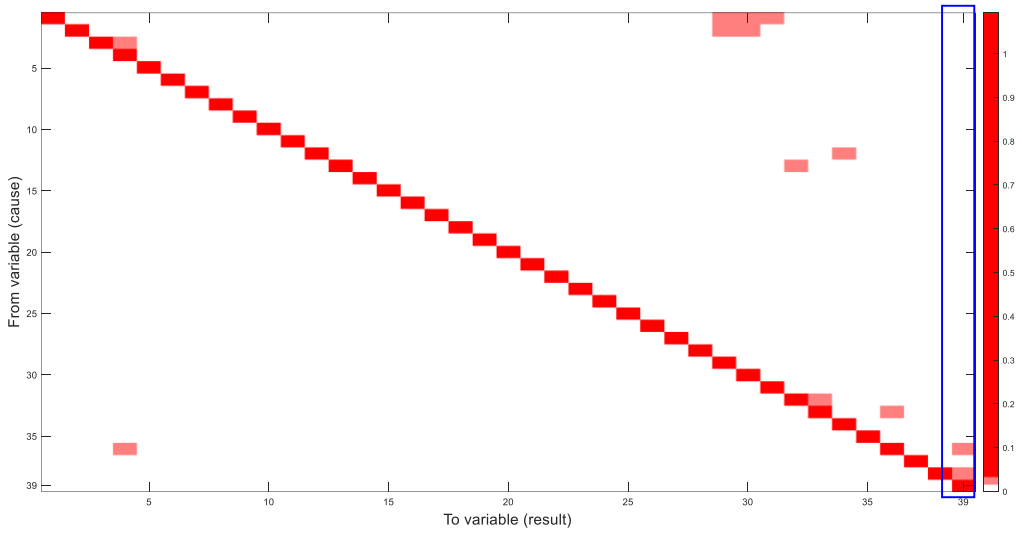

Figure 14. TE analysis for SFA

To compare the inputs number of different inferential sensors, table 3 lists the number of inputs used in TE latent causal feature learning algorithm. Table 4 lists the number of inputs used in TE causal feature learning algorithm.

Table 3. Number of inputs in TE latent causal feature learning

Inferential sensor	Number of inputs
PCA+TE	**1**
SFA+TE	**2**
variable+TE	**28**

Table 4. Number of inputs in TE causal feature learning

Inferential sensor	Number of inputs
TE+PCA	**9**
TE+SFA	**13**
variable+TE	**28**

4.2. Results of Inferential Sensor based on Causality Analysis

A. Comparison between Causality Analysis and Regularization

Fig15 shows the results of dynamic least square regression (DLSR), elastic net, DLSR with TE feature selection, and DLSR with GCA feature selection. The inputs number of these inferential sensors are 66, 17, 28, 13 respectively. The left side is the training process and the right side is test

process. Table 5 lists the results of different inferential sensors.

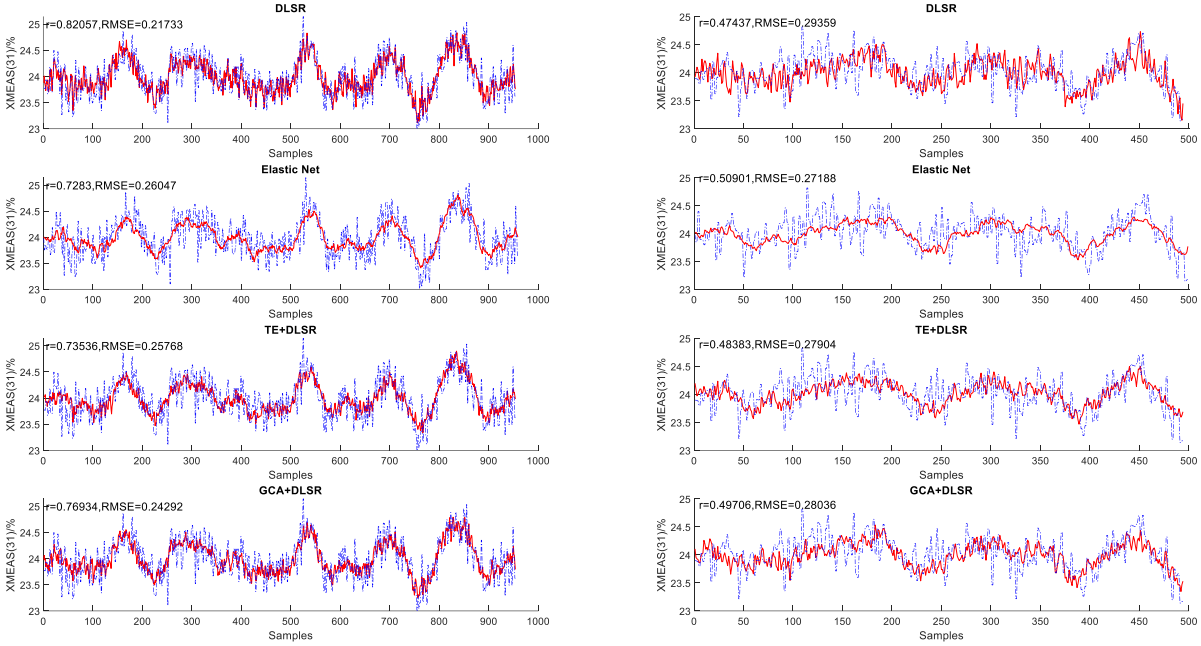

Figure 15. Results of causality analysis and noncausality analysis in inferential sensors

Table 5. Performance of causal and noncausal inferential sensors

Inferential	r_{train}	$RMSE_{train}$	r_{test}	$RMSE_{test}$
DLSR	0.82057	0.21733	0.47437	0.29359
Elastic Net	0.7283	0.26047	**0.50901**	**0.27188**
TE+DLSR	0.73536	0.25768	0.48383	0.27904
GCA+DLSR	0.76934	0.24292	0.49706	0.28036

B. Inferential Sensor based on GCA

In this part, causal feature learning and latent causal feature learning based on GCA are analyzed. For causal feature learning inferential sensors, we first select causal features, and then extract latent features. Fig 16-17 show the results of causal feature learning in training process and test process, respectively. Table 6 lists the results of causal feature learning inferential sensors based on GCA. In training process, as was expected, $RMSE_{train}$ of inferential sensors without causal feature learning is lower and r_{train} is higher. However, in test process, the performance of causal feature learning is better.

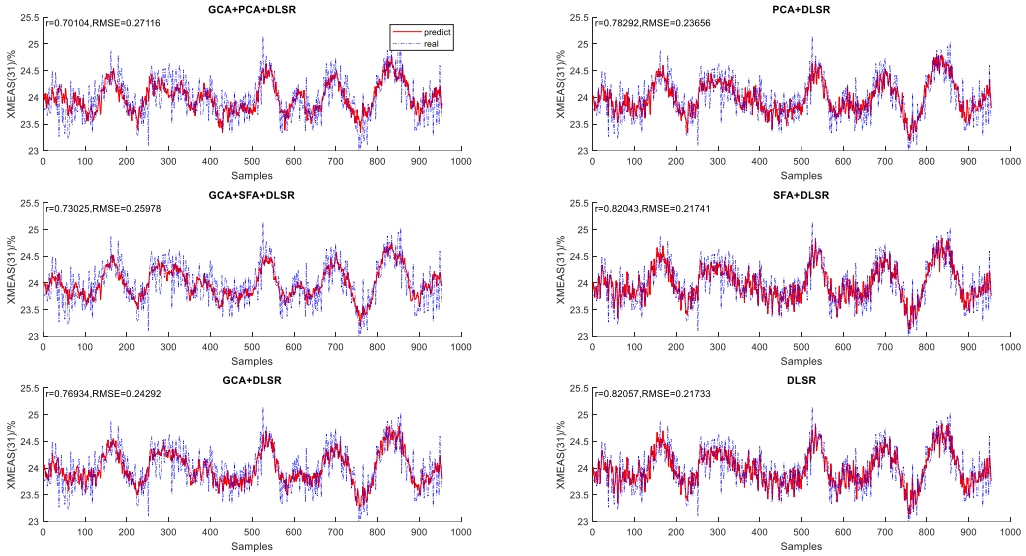

Figure 16. Results of GCA causal feature learning inferential sensors in training process

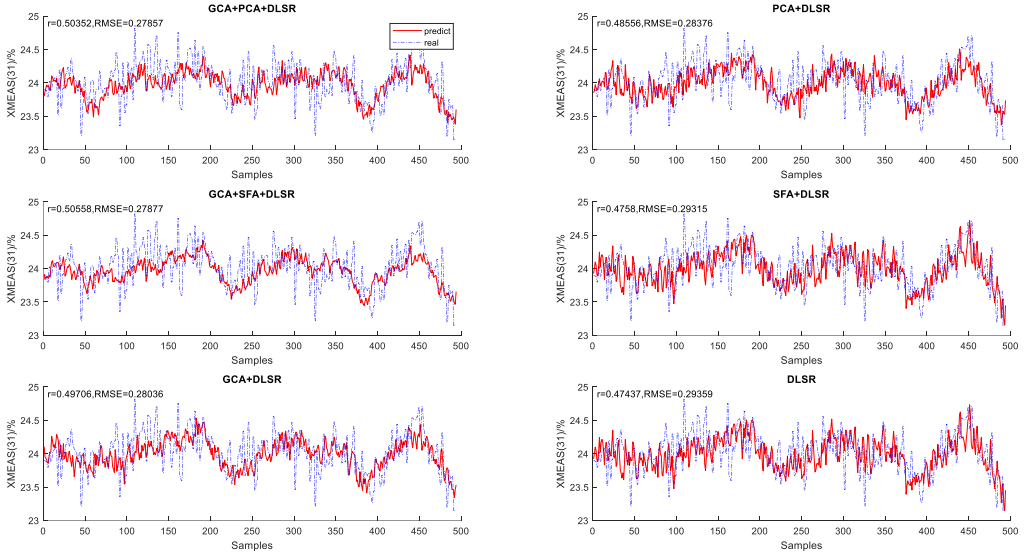

Figure 17. Results of GCA causal feature learning inferential sensors in test process

Table 6. Performance of GCA causal feature learning inferential sensors

Inferential Sensor	r_{train}	$RMSE_{train}$	r_{test}	$RMSE_{test}$
DLSR	0.82057	0.21733	0.47437	0.29359
GCA+DLSR	0.76934	0.24292	0.49706	0.28036
GCA+PCA+DLSR	0.70104	0.27116	**0.50352**	**0.27857**
PCA+DLSR	0.78292	0.23656	0.48556	0.28376
GCA+SFA+DLSR	0.73025	0.25978	**0.50558**	**0.27877**
SFA+DLSR	0.82043	0.21741	0.4758	0.29315

For latent causal feature learning inferential sensors, we first extract latent features and then select causal features. Fig 18-19 show the results of training process and test process, respectively. Table 7 lists the results of latent causal feature learning inferential sensors based on GCA.

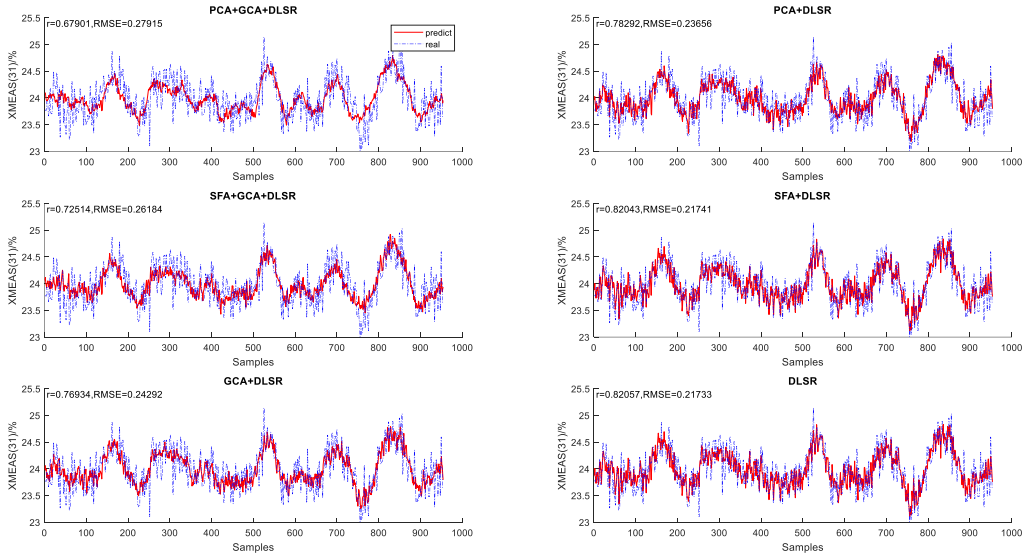

Figure 18. Results of GCA latent causal feature learning inferential sensors in training process

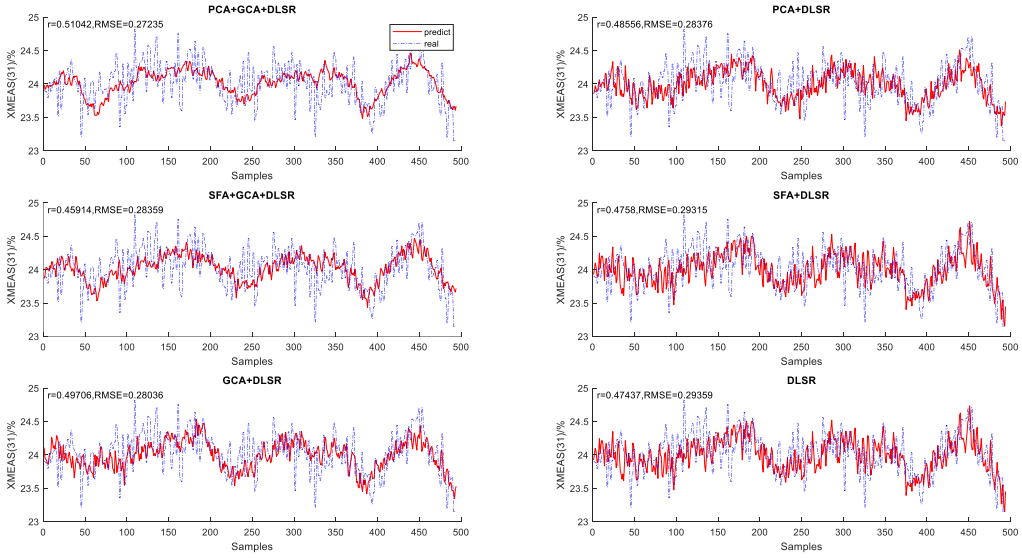

Figure 19. Results of GCA latent causal feature learning inferential sensors in test process

Table 7. Performance of GCA latent causal feature learning inferential sensors

Inferential Sensor	r_{train}	$RMSE_{train}$	r_{test}	$RMSE_{test}$
DLSR	0.82057	0.21733	0.47437	0.29359
GCA+DLSR	0.76934	0.24292	0.49706	0.28036

PCA+GCA+DLSR	0.67901	0.27915	**0.51042**	**0.27235**
PCA+DLSR	0.78292	0.23656	0.48556	0.28376
SFA+GCA+DLSR	0.72514	0.26184	**0.45914**	**0.28359**
SFA+DLSR	0.82043	0.21741	0.4758	0.29315

C. Inferential Sensor based on TE

In this part, causal feature learning and latent causal feature learning inferential sensors based on TE are analyzed. Fig 20-21 show the results of causal feature learning in training process and test process, respectively. Table 8 lists the results of causal feature learning based on TE. Fig 22-23 show the results of latent causal feature learning in training process and test process, respectively. Table 9 lists the results of latent causal feature learning based on TE.

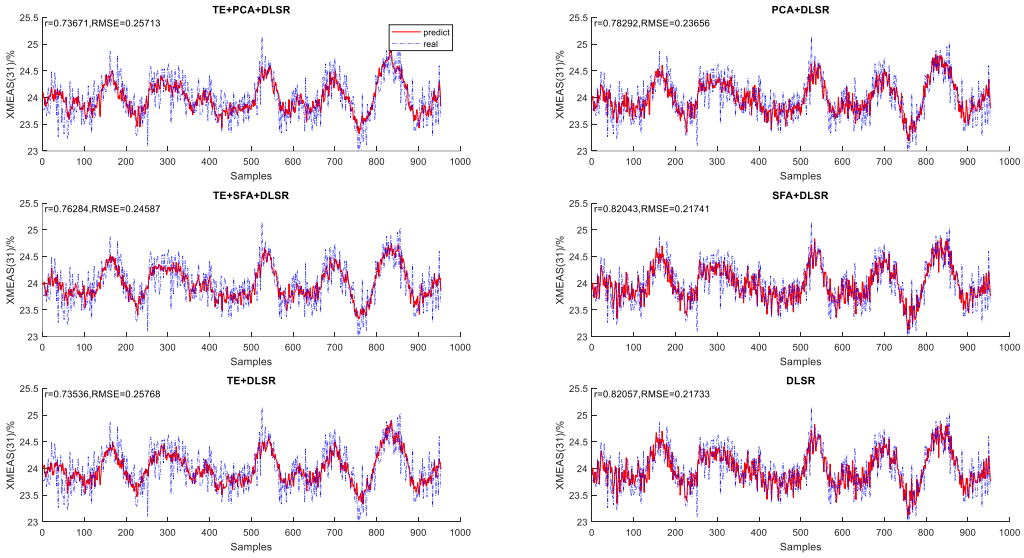

Figure 20. Results of TE causal feature learning inferential sensors in training process

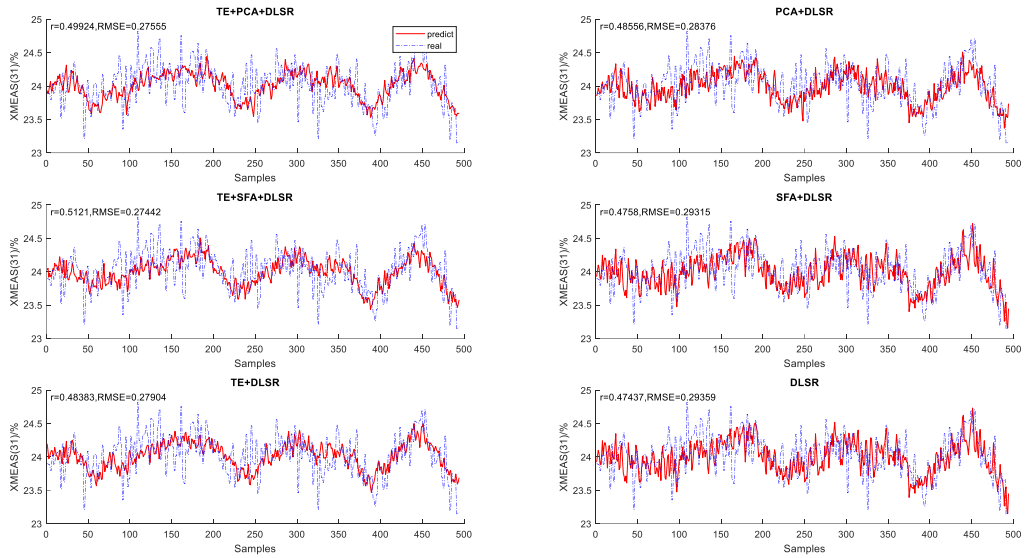

Figure 21. Results of TE causal feature learning inferential sensors in test process

Table 8. Performance of TE causal feature learning inferential sensors

Inferential Sensor	r_{train}	$RMSE_{train}$	r_{test}	$RMSE_{test}$
DLSR	0.82057	0.21733	0.47437	0.29359
TE+DLSR	0.73536	0.25768	0.48383	0.27904
TE+PCA+DLSR	0.73671	0.25713	**0.49924**	**0.27555**
PCA+DLSR	0.78292	0.23656	0.48556	0.28376
TE+SFA+DLSR	0.76284	0.24587	**0.5121**	**0.27442**
SFA+DLSR	0.82043	0.21741	0.4758	0.29315

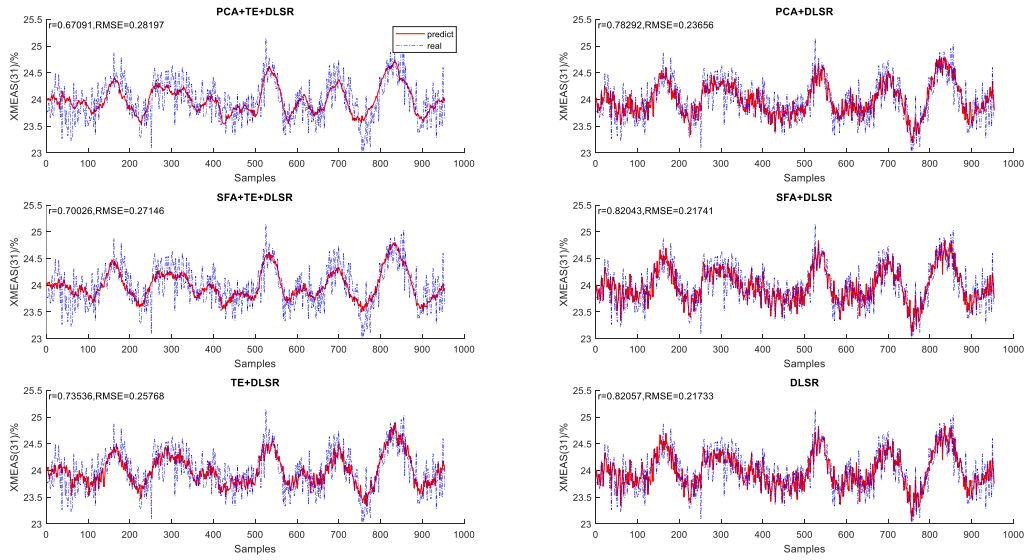

Figure 22. Results of TE latent causal feature learning inferential sensors in training process

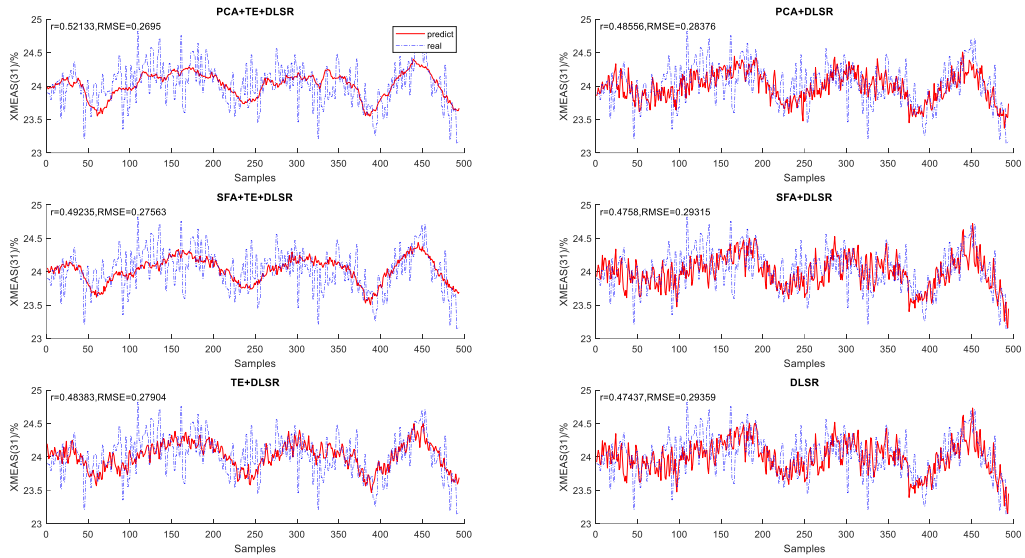

Figure 23. Results of TE latent causal feature learning inferential sensors in test process

Table 9. Performance of TE latent causal feature learning inferential sensors

Inferential Sensor	r_{train}	$RMSE_{train}$	r_{test}	$RMSE_{test}$
DLSR	0.82057	0.21733	0.47437	0.29359
TE+DLSR	0.73536	0.25768	0.48383	0.27904
PCA+ TE+DLSR	0.67091	0.28197	**0.52133**	**0.2695**
PCA+DLSR	0.78292	0.23656	0.48556	0.28376
SFA+ TE+DLSR	0.70026	0.27146	**0.49235**	**0.27563**
SFA+DLSR	0.82043	0.21741	0.4758	0.29315

Similar with the results of GCA, it can be easily found that TE combining with the latent variable models has the best performance with minimal inputs compared with other inferential sensors.

5. Discussions

Table 10 lists detailed comparison results of different inferential sensors. The number in bold blue means the best performance in all algorithms while the bold red means the worst performance. Generally, the best performance of only using causality analysis (GCA, TE) is RMSE:0.27904 and r: 0.49706; the best performance of only using latent variable model (PCA, SFA) is RMSE:0.28376 and r: 0.48556; However, almost all performance of latent causal feature learning or causal feature learning are better than only using causality analysis or latent variable model, it is obvious that the performance is improved when combining them.

In addition, the maximum input of latent causal feature learning is 6 and maximum input of causal feature learning is 13, the minimum input of causality analysis is 13 and the minimum inputs

of latent variable model is 24. These results show that proposed algorithms have better performance with fewer inputs, which proves that the integration of causality analysis and latent variable models is powerful in simplifying model without losing important information.

It can be seen that latent causal feature learning based on PCA+TE has the best results, which, only using 1 latent causal feature, achieves the minimal $RMSE_{test}$ and maximum r_{test}. Compared with the results of elastic net, original variables, the results of causality analysis show that the model complexity is reduced. Compared with the results of latent variable model and causality analysis, it shows that there exist some unimportant features in both methods; the combination of causality analysis and latent variable model makes it possible to find which features are important for building inferential sensors.

Table 10. Comparison of different inferential sensors

Performance index		r_{test}		$RMSE_{test}$		Input numbers	
Inferential sensors		GCA	TE	GCA	TE	GCA	TE
Latent causal feature learning	PCA	0.51042	**0.52133**	0.27235	**0.2695**	3	**1**
	SFA	**0.45914**	0.49235	**0.28359**	0.27563	**6**	2
Causal feature learning	PCA	0.50352	**0.49924**	0.27857	0.27555	9	**9**
	SFA	0.50558	**0.5121**	**0.27877**	**0.27442**	**7**	13
Causality analysis		0.49706	**0.48383**	0.28306	0.27904	13	28
Latent variable mode	PCA	**0.48556**		**0.28376**		**24**	
	SFA	**0.4758**		**0.29315**		**38**	
Elastic net		**0.50901**		**0.27188**		**17**	
Original Variables		**0.47437**		**0.29359**		**66**	

Based on the previous discussion, unlike traditional methods (Latent variable model, Elastic net) which either cannot provide precise prediction using simple model or suffer from high model complexity, the advantages of proposed dynamic inferential sensor based on causality analysis are as follows:

A. Inferential sensor based on causality analysis is more robust since causal features reflect underlying dynamics of the process data, while original independent variables contain lots of noise information. The introduction of causality analysis can select important features and achieve noise reduction.

B. There are many feature-extraction methods in industrial processes, making it hard to tell which types of features are important. However, the introduction of causality analysis making it possible to find important features.

C. The inputs of inferential sensor are causal variables, and these input causal variables have

causality relationship with the outputs, Therefore, the interpretability of proposed inferential sensor is enhanced.

D. The integration of causality analysis and latent variable models is very useful, since there is no correlation between each latent feature in most latent variable models like SFA and PCA. This advantage is very useful in reducing the negative impact of correlation and simplifying the inferential sensor model.

6. Conclusion

Using the idea of causality analysis, two types of dynamic feature learning schemes based on causality analysis are proposed to develop inferential sensors. With dynamic latent causal features obtained from causal feature learning or latent causal feature learning, the proposed inferential sensors are more robust since causal features reflect underlying dynamics of the process data. The interpretability of inferential sensor is enhanced with causal feature selection and latent feature extraction. Case studies demonstrate that the complexity of inferential sensor is sharply decreased while the prediction performance is improved. This study has shown that integration of causality analysis and latent variable model is promising in designing inferential sensors.

References

[1] C. Shang (2018). *Dynamic modeling of complex industrial processes: Data-driven methods and application research*, Springer.

[2] S. X. Ding, S Yin, K. X. Peng, H. Hao and B. Shen (2013). A Novel Scheme for Key Performance Indicator Prediction and Diagnosis with Application to an Industrial Hot Strip Mill. *IEEE Transactions on Industrial Informatics*, 9(4), 2239-2247.

[3] S Yin, H. J. Gao and O. Kaynak (2014). Data-Driven Control and Process Monitoring for Industrial Applications—Part I. *IEEE Transactions on Industrial Electronics*, 61(11), 6356-6359.

[4] C. Shang, X. Gao, F. Yang and D. X. Huang (2014). Novel Bayesian framework for dynamic soft sensor based on support vector machine with finite impulse response. *IEEE Transactions on Control Systems Technology*, 22(4), 1550-1557.

[5] M. Kano *et al.* (2000). Inferential control system of distillation compositions using dynamic partial least squares regression. *Journal of Process Control*, 10, 157–166.

[6] W.X. Lu *et al*. (2008). A Dynamic Soft-sensing Method Based on Impulses Response Template and Parameter Estimation with Modified DE Optimization. *IFAC Proceedings Volumes*, 41(2),10983-10988.

[7] X. Yuan, L. Li and Y. Wang (2020). Nonlinear dynamic soft sensor modeling with supervised long short-term memory network. *IEEE Transactions on Industrial Informatics*, 16(5), 3168-3176.

[8] Y. N. Dong, S. J. Qin (2018). Dynamic latent variable analytics for process operations and control. *Computers & Chemical Engineering*, 114, 69-80.

[9] J. Federico, J. Michael and P. Tomaso (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7(2), 219-269.

[10] Y. Liu and M. Schumann (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, 56(9), 1099-1108.

[11] L. Barnett, A. B. Barrett and A. K. Seth (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters*, 103(23),238701.

[12] C. W. J. Granger (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424–438.

[13] W. Hu, J. Wang, T. Chen and Shah SL (2017). Cause-effect analysis of industrial alarm variables using transfer entropies. *Control Engineering Practice*, 64, 205-214.

[14] W. Yu, F. Yang (2015). Detection of causality between process variables based on industrial alarm data using transfer entropy. *Entropy*, 17(8), 5868-5887.

[15] H. Gharahbagheri, S. Imtiaz A and F. Khan (2017). Root cause diagnosis of process fault using KPCA and Bayesian network. *Industrial & Engineering Chemistry Research*, 56(8), 2054-2070.

[16] T. Schreiber (2000). Measuring information transfer. *Physical Review Letters*, 85, 461–464.

[17] Q. X. Zhu, Y. Luo and Y. L. He (2019). Novel Multiblock Transfer Entropy Based Bayesian Network and Its Application to Root Cause Analysis. *Industrial & Engineering Chemistry Research*, 58(12), 4936-4945.

[18] L. Rinat (2019). Data-based causality analysis by exploiting process connectivity information. Ph.D. dissertation, Aalto University.

[19] W. Hu, J. D. Wang, T. W. Chen and S. L. Shah (2017). Cause-effect analysis of industrial alarm variables using transfer entropies. *Control Engineering Practice*, 64, 205-214.

[20] P. Duan, F. Yang, S. L. Shah and T. W. Chen (2015). Transfer zero-entropy and its application for capturing cause and effect relationship between variables. *IEEE Transactions on Control Systems Technology*. 23(3),855–867.

[21] R. Xie, N. M. Jan, K. Hao, L. Chen and B. Huang (2020). Supervised Variational Autoencoders for Soft Sensor Modeling with Missing Data. *IEEE Transactions on Industrial Informatics*, 16(4), 2820-2828.

[22] X. Yuan, B. Huang, Y. Wang, C. Yang and W. Gui (2018). Deep Learning-Based Feature Representation and Its Application for Soft Sensor Modeling with Variable-Wise Weighted SAE. *IEEE Transactions on Industrial Informatics*, 14(7), 3235-3243.

[23] S. Valle, W. H. Li, S. J Qin (1999). Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Industrial & Engineering Chemistry Research*, 38(11), 653-658.

[24] Z. Zhang et al. (2017). Discriminative Elastic-Net regularized linear regression. *IEEE Transactions on Image Processing*, 26(3), 1466-1481.

[25] C. Shang, F.Q. You (2018). Robust optimization in high-dimensional data space with support vector clustering. *IFAC-Papers Online*, 51(18), 19-24.

[26] L. H. Chiang, E. L. Russell, R.D. Braatz (2000). Tennessee Eastman Process. In: *Fault Detection and Diagnosis in Industrial Systems*. London, UK: Springer, 103-112.