4th International Conference on Industrial Artificial Intelligence (IAI), Aug. 24-27, 2022, Shenyang, China

Sparse Causality Analysis Approach with Time-varying Parameters for Root Cause Localization of Nonstationary Process

Pengyu Song, Chunhui Zhao, *Senior Member*, *IEEE*, Biao Huang, *Fellow*, *IEEE*, Jinliang Ding, *Senior Member*, *IEEE*

***Abstract*—Root cause diagnosis (RCD) is an important technique for maintaining the safe operation of industrial processes. Traditional RCD methods usually require stationarity assumptions. However, the process inevitably shows nonstationarity due to factors such as switching of operating conditions. Although there have been some previous studies trying to overcome the challenge of nonstationarity, these methods fail to guarantee the significance of the extracted causalities and lead to redundant relationships. To address the above issues, a sparse causal analysis model with time-varying parameters is extracted in this study. First, we propose an end-to-end information fusion and prediction task to characterize predictive relationships between variables and avoid repeated modeling. Second, we design time-varying parameters for the information fusion mechanism to cope with nonstationarity and automatically identify significant causality through sparse parameter updates. We design an update strategy that constrains the gradient information to guarantee sparsity. Finally, a causal metric is constructed for the time-varying predictive relationship to comprehensively obtain the overall causal relationship, which further guarantees causal significance. The validity of the proposed method is illustrated through a real industrial example collected from a thermal power plant.**

Keywords—time-varying parameter updating, nonstationary industrial process, root cause localization, fault diagnosis, sparsity constraint

I. Introduction

Generally, an unsupervised intelligent maintenance model of an industrial process can be composed of three parts, including fault detection, isolation, and diagnosis. Fault detection models are designed to determine in real-time whether operations are in a healthy state, automatically identifying process faults. After a fault is detected, an isolation model is activated to identify the key fault variables that constitute the set of candidates that are the root cause of the fault. Subsequently, the root cause diagnosis (RCD) model was used to mine the causal relationships between the faulty variables, thereby identifying the root cause variables of the fault. Although fault detection [1]-[4] and isolation models [5], [6] have been well explored, research on RCD is just emerging. Classical industrial process RCD methods can be mainly divided into three categories, Granger causality (GC), transfer entropy (TE), and dynamic Bayesian networks (DBN).

GC [7] is one of the most classic time series causal inference methods, which characterize causality through predictive relationships among variables. Specifically, if the introduction of past information on one variable $\mathbf{x}$ can significantly improve the prediction accuracy of another variable $\mathbf{y}$, then $\mathbf{x}$ is considered to be the Granger cause of $\mathbf{y}$. In the most primitive GC models, prediction (AR) is achieved by autoregression. However, AR is a univariate linear model and cannot handle the general nonlinearity and multivariate interactions. To solve nonlinear problems, many advanced regression models are used to replace AR in GC models. Chen et al. [8] proposed the GPR-Granger method, which uses Gaussian process regression (GPR) [9] as a predictive model to account for nonlinearity. In addition, neural networks are also used for temporal prediction in GC. To consider the interaction of multiple variables, two types of methods have been proposed successively, including sparse variable selection and vector autoregressive (VAR) models. The former uses the sampling of all variables at each time lag as the input of the prediction model and selects the most useful variables for prediction through sparse constraints, thereby extracting significant causal relationships in multivariate scenarios. Representative methods include Lasso Granger (LG) [10], etc. The latter uses VAR for multivariate regression to directly extract multivariate predictive relationships. Some frequency domain indicators, such as partial directed coherence (PDC) [11], are constructed to express causality in VAR Granger.

TE [12]-[14] is a causal inference method based on information theory. It uses entropy to measure the temporal information transfer between variables to determine cause and effect. When the variables follow Gaussian distributions, TE is verified to be equivalent to GC [15]. Furthermore, since TE uses probability distributions to measure information transfer, it can directly handle data with nonlinearities.

DBN [16] is a probabilistic graphical model that describes the dependencies between variables in a probabilistic framework. The research on DBN is divided into two aspects, structure learning and parameter learning. The former aims to find the best graph structure that can describe the relationship between variables, while the latter focuses on learning the

*This work was supported in part by the National Key R&D Program of China (2021YFC2902701), the National Natural Science Foundation of China (No. 62125306), the Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (ICT2021A15), and State Key Laboratory of Synthetical Automation for Process Industries (2020-KF-21-07). *(Corresponding author: Chunhui Zhao)*

Pengyu Song and Chunhui Zhao are with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China. Email: pysong@zju.edu.cn (P. Song), chhzhao@zju.edu.cn (C. Zhao).

Biao Huang is with the Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB T6G 2G6, Canada (e-mail: biao.huang@ualberta.ca).

Jinliang Ding is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China (e-mail: jlding@mail.neu.edu.cn)

2022 4th International Conference on Industrial Artificial Intelligence (IAI) | 978-1-6654-5120-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IAI55780.2022.9976691

978-1-6654-5120-8/22/$31.00©2022 IEEE

parameters in the graph structure under the premise of sufficient historical events. Since structural learning usually requires complex heuristic searches and requires large amounts of data [17], parameter learning of DBNs is studied in the process industry in the presence of sufficient historical failure events [18], [19]. However, this study focuses on how to determine the root cause of an unseen failure event without historical data. Therefore, we will not discuss DBNs much in the following.

Although TE and GC methods have been widely studied in industrial RCD tasks, they require the process data to be stationary. However, the process often exhibits nonstationary characteristics due to factors such as switching of operating conditions. In addition, faults often lead to abnormal fluctuations in the process, which makes nonstationary characteristics more common in fault data. How to overcome nonstationarity is an urgent challenge for RCD tasks that require the use of fault data.

For GC methods, difference and time-varying methods are the main ways to resolve nonstationarity. The difference method converts the original nonstationary data into a stationary sequence through several difference operations and then conducts causal inference. However, this approach causes unavoidable loss of information and cannot account for long-term dependencies. A previous study [20] proposed a vector error correction model (VECM)-based GC method that comprehensively considered causality in raw and differential data. However, this method requires the data to be integrated of order one. Time-varying methods are another idea for constructing nonstationary GC models, which allow the parameters in the model to vary over time to adapt to the time-varying predictive relationships caused by nonstationarity. This type of method does not require any local stationarity assumption and does not cause information loss, so it has better practicality. Typical parameter updating algorithms in time-varying models include least mean squares (LMS) [21], recursive least squares (RLS) [22], variational Bayes (VB) [23],and Kalman filter [24]

For TE methods, symbolization methods can be used to tackle nonstationarity. By symbolizing the original time series as local rank vectors, TE methods can capture local patterns without the need for non-stationarity assumptions. Currently, symbolic TE (STE) [25] and its multivariate version, partial symbolic TE (PSTE) [26], are proposed and applied to RCD [27] in industrial processes. However, symbolization can also cause loss of information like differencing operations.

Although the above methods have been used to solve nonstationary problems, none of them can guarantee the significance of the extracted causality. Specifically, it has been shown that sparsity constraints should be introduced in causality extraction to select key causalities [28]. Otherwise, too many redundant relations will affect the inference accuracy. Considering the advantages of time-varying models over differencing and symbolic methods, we aim to build a time-varying GC model with sparsity constraints, which can handle nonstationarity while guaranteeing causal significance. However, this leads to a theoretical challenges, that is, commonly used sparsity constraints, such as Lasso [29], elastic net (EN) [30], etc., cannot be directly applied in time-varying models to induce sparsity [31].

In this work, we propose a novel GC method with sparse and time-varying parameters and apply it to the RCD task of nonstationary industrial processes. First, we design an end-to-end information fusion and prediction mechanism, which can capture multivariate predictive relations without repeated modeling in time-varying scenes. Second, a parameter update algorithm with sparse constraints is proposed to guarantee causal saliency. The sparsity of the parameters can be guaranteed mathematically. Finally, a causal metric and root cause scoring mechanism is established, which can extract causalities characterizing fault propagation paths from time-varying predictive relationships and quantify root cause scores for each variable. The main contributions are summarized as follows:

1) The proposed time-varying GC method can obtain sparse solutions to overcome the nonstationary problem and ensuring causal significance, reducing redundant causal relationships to improve diagnostic performance.

2) A causal metric and root cause scoring mechanism for time-varying models is designed to provide quantitative root cause representations.

The rest of this paper is structured as follows. The LG method is revisited in Section II. In Section III, we demonstrate the proposed method in detail. Afterward, the performance of the proposed RCD method is verified by experimental results in Section IV. Finally, conclusions are drawn in Section V.

II. Revisit of Lasso Granger

GC can be characterized as a significant predictive contribution. Traditional GC is only suitable for bivariate scenarios, and the significance of predictive relationships is determined by hypothesis testing. Therefore, it cannot handle the interaction between variables well in multivariate scenarios. LG [10] can solve this problem through the Lasso sparse constraint [29]. For N variables $\left\{\mathbf{x}_1, \mathbf{x}_{2,} \ldots, \mathbf{x}_N\right\}$, LG builds the following forecasting model for each variable:

$$\hat{x}_i\left(t\right) = \sum_{j=1}^{N} \mathbf{w}_{ji}^{\mathrm{T}} \mathbf{x}_{j,[t-T,t-1]} \qquad (1)$$

where T is the time lag; $\mathbf{x}_{j,[t-T,t-1]} = \left[x_j\left(t-T\right), \ldots, x_j\left(t-1\right)\right]^{\mathrm{T}}$, denoting the lagged sampling vector of variable $\mathbf{x}_i$; $\mathbf{w}_{ji} \in \Re^T$, which is the predictive vector from $\mathbf{x}_j$ to $\mathbf{x}_i$. The optimization objective of LG is:

$$\min_{\mathbf{w}_{ji}} \left[x_i\left(t\right) - \hat{x}_i\left(t\right)\right]^2 + \alpha \sum_{j=1}^{N} \left\|\mathbf{w}_{ji}\right\|_1 \qquad (2)$$

where α is the L1 penalty factor.

The Lasso method i.e., the L1 norm penalty term, can make the vector $\mathbf{w}_{ji}$ show sparsity. By minimizing the prediction error and L1 norm, only the variables that have significant prediction contributions can have nonzero coefficients in $\mathbf{w}_{ji}$. Thus, the significant predictive relations are selected to reflect GC. Specifically, if a the vector $\mathbf{w}_{ji}$ become a zero vector after optimization, it means that the information contained in $\mathbf{x}_j$ is not useful for the forecasting of $\mathbf{x}_i$ at all time lags, so $\mathbf{x}_j$ is not the Granger cause of $\mathbf{x}_i$. Otherwise, $\mathbf{x}_j$ is determined to be the Granger cause of $\mathbf{x}_i$.

Since LG builds a predictive model with multiple variables together, it can take multivariate interactions into account. However, LG still models each variable once and requires stationarity assumptions.

III. Methodology

In this section, the proposed time-varying prediction model and parameter updating algorithm are presented first. Afterward, we provide the designed causal metric, which aims to obtain overall causalities for the time-varying model.

A. Sparse and time-varying predictive relation extraction

We give a detailed description of the proposed method in this subsection. First, an end-to-end information fusion and prediction model for a multivariate process is provided, which is the basis of the entire model. Next, we extend the base model to a time-varying version and give its parameter update strategy. In the designed strategy, parameters of different statuses can have different update modes so as to ensure the rationality of the extracted predictive relations in time-varying scenarios.

1) End-to-end information fusion and prediction model

The LG model introduced above extracts multivariate causality by selecting key predictive relationships through the Lasso penalty term. However, LG still needs to model each variable once. Since time-varying models require model updates at each sample point, we hope to make end-to-end predictions without repeated modeling. To this end, we provide a straightforward multivariate information fusion and prediction model and extend it to a time-varying version to handle nonstationarity.

Inspired by the previous study [32], for a multivariate time series $\mathbf{X}$ with N variables ($\mathbf{x}_i$, $i=1,\dots,N$), we use an information fusion matrix $\mathbf{C}$ to represent the predictive dependencies among variables in the proposed model, where $\mathbf{C}\in\Re^{N\times N}$. The element in the *i*-th row and the *j*-th column of $\mathbf{C}$ matrix is denoted as c_{ij}, which is used to describe the predictive relation from $\mathbf{x}_j$ to $\mathbf{x}_i$. In the prediction process, the information of each variable is fused according to the $\mathbf{C}$ matrix. Mathematically, the prediction model can be formulated as:

$$\hat{x}_i(t)=\sum_{j=1}^{N}c_{ij}\mathbf{x}_{j,[t-T:t-1]}\mathbf{w}_i\,(t>T) \tag{4}$$

where $\mathbf{w}_i$ is a the linear mapping associated with $\mathbf{x}_i$, which is used to extract temporal features; $\mathbf{x}_{j,[t-T:t-1]}$ is the lagged sampling vector of $\mathbf{x}_j$ consisting of samples of $\mathbf{x}_j$ from the ($t-T$)-th to the ($t-1$)-th sampling point; T is the time lag; $\hat{x}_i(t)$ is the predicted result of $x_i(t)$.

The optimization objective is to minimize the prediction error of each variable, which can be calculated as:

$$\min_{c_{ij},\mathbf{w}_i}\left\|x_i(t)-\hat{x}_i(t)\right\|_2^2+\alpha\sum_{j=1}^{N}\left|c_{ij}\right|+\beta\sum_{j=1}^{N}c_{ij}^2 \tag{5}$$

where α and β are the penalty coefficients of L1 and L2 norms, respectively. They are also set to be positive numbers.

In the prediction model presented above, the element c_{ij} controls the information fusion between variable $\mathbf{x}_i$ and $\mathbf{x}_j$. If the value of c_{ij} is in a relatively large order of magnitude, the information of $\mathbf{x}_j$ will affect the prediction result of $\mathbf{x}_i$ more significantly, and vice versa. In a special case, if the value of c_{ij} is 0, it means that the prediction of $\mathbf{x}_i$ does not depend on $\mathbf{x}_j$ at all. Besides, it is worth noting that we add a sparsity penalty for c_{ij}, i.e., the elastic net (EN) constraint [30] to the optimization objective. EN combines L1 and L2 norm constraints to achieve parameter sparsification, which is shown to have better variable selection performance than Lasso. In this way, while reducing prediction error, relationships that are more useful for prediction are preserved and redundant relationships are reduced to 0, thereby capturing significant causality. The absolute value of c_{ij} ($\left|c_{ij}\right|$) can represent the prediction contribution from $\mathbf{x}_j$ to $\mathbf{x}_i$, indicating GC between variables. A larger $\left|c_{ij}\right|$ means $\mathbf{x}_j$ has a more significant GC to $\mathbf{x}_i$.

Moreover, the prediction model in Eq. (4) can be converted to a matrix form, which can be expressed as:

$$\hat{\mathbf{x}}(t)=\operatorname{diag}\left(\mathbf{C}\mathbf{X}_{[t-T:t-1]}\mathbf{W}\right)(t>T) \tag{6}$$

where $\mathbf{X}_{[t-T:t-1]}$ is matrix composed of the lagged sampling vector of all variables; $\operatorname{diag}(\cdot)$ means taking the diagonal elements of the matrix; $\mathbf{W}=\left[\mathbf{w}_1,\dots,\mathbf{w}_N\right]$, namely the linear mapping matrix constructed from the linear mapping vectors of each variable.

Accordingly, the optimization objective function in the matrix form can be written as:

$$\min_{\mathbf{C},\mathbf{W}}\left\|\mathbf{x}(t)-\hat{\mathbf{x}}(t)\right\|_2^2+\alpha\sum_{i=1}^{N}\sum_{j=1}^{N}\left|c_{ij}\right|+\beta\sum_{i=1}^{N}\sum_{j=1}^{N}c_{ij}^2 \tag{7}$$

The matrix form suggests that we can perform an end-to-end calculation and optimization on the proposed information fusion and prediction model, which avoids repeated modeling for each variable.

In addition, the above prediction model should be extended to a time-varying version:

$$\hat{\mathbf{x}}(t)=\operatorname{diag}\left[\mathbf{C}(t)\mathbf{X}_{[t-T:t-1]}\mathbf{W}(t)\right](t>T) \tag{8}$$

and the optimization objective is:

$$\min_{\mathbf{C}(t),\mathbf{W}(t)}\left\|\mathbf{x}(t)-\hat{\mathbf{x}}(t)\right\|_2^2+\alpha\sum_{i=1}^{N}\sum_{j=1}^{N}\left|c_{ij}(t)\right|+\beta\sum_{i=1}^{N}\sum_{j=1}^{N}c_{ij}^2(t) \tag{9}$$

where $\mathbf{C}(t)$ and $\mathbf{W}(t)$ are the information fusion matrix and the linear mapping matrix at the *t*-th time point, respectively.

2) Parameter updating strategy

The time-varying mechanism allows the model to have different parameters at different sample points to cope with time-dependent predictive relationships. However, none of the existing time-varying RCD methods can introduce sparsity to guarantee causal saliency. In this subsection, we show a parameter update strategy that can lead to sparse solutions.

Before giving the designed model update strategy, the classical time-varying parameter estimation model, LMS [21], is formulated as follows:

$$\vartheta(t+1)=\vartheta(t)-\eta\frac{\partial Loss}{\partial\vartheta(t)} \tag{10}$$

where $\vartheta(t)$ denotes the model parameter at the *t*-th sample, *Loss* is the loss function of the optimization objective function, and η is the learning rate.

LMS can adaptively update model parameters through gradient information at each moment, but it cannot obtain sparse solutions even when sparse constraints such as L1 norm or EN are introduced. This is because the sparsification constraint contains non-smooth terms, which is non-differentiable at zero, and LMS can hardly converge to the non-differentiable point with only gradient information [31].

In the proposed method, there are two types of parameters should be updated with time, including the information fusion matrix $\mathbf{C}(t)$ and the linear mapping matrix $\mathbf{W}(t)$. For $\mathbf{W}(t)$, we can be found that there is no non-smooth penalty on it. Therefore, the LMS algorithm can be directly applied to $\mathbf{W}(t)$, namely:

$$\mathbf{W}(t+1)=\mathbf{W}(t)-\eta\frac{\partial L}{\partial\mathbf{W}(t)} \tag{11}$$

where *L* is the loss function in Eq. (9), which can be expressed as:

$$\begin{aligned}L=&\left\|\mathbf{x}(t)-\operatorname{diag}\left[\mathbf{C}(t)\mathbf{X}_{[t-T:t-1]}\mathbf{W}(t)\right]\right\|_2^2\\&+\alpha\sum_{i=1}^{N}\sum_{j=1}^{N}\left|c_{ij}(t)\right|+\beta\sum_{i=1}^{N}\sum_{j=1}^{N}c_{ij}^2(t)\end{aligned} \tag{12}$$

and the gradient can be calculated as:

$$\left[\frac{\partial L}{\partial\mathbf{W}(t)}\right]_i=2\left(\hat{x}_i(t)-x_i(t)\right)\left[\mathbf{X}_{[t-T:t-1]}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}(t)\right]_i \tag{13}$$

where $[\cdot]_i$ means taking the *i*-th column of the matrix.

Next, we provide the updating method for $\mathbf{C}(t)$. Considering the first term in *L*, i.e., the prediction error (denoted as *P*), contains the interactive operation of different elements in the matrix $\mathbf{C}(t)$, we perform a first-order Taylor expansion on *P* to simplify:

$$\begin{aligned}&P\left(\mathbf{C}^*(t)\right)=P\left(\mathbf{C}(t)+\Delta\mathbf{C}(t)\right)\\&\approx P\left(\mathbf{C}(t)\right)+\operatorname{vec}^{\mathrm{T}}\left(\frac{\partial P}{\partial\mathbf{C}(t)}\right)\operatorname{vec}\left(\mathbf{C}^*(t)-\mathbf{C}(t)\right)\end{aligned} \tag{14}$$

Substituting the above simplified result into the optimization objective in Eq. (9), the original problem is transformed into:

$$c_{ij}(t+1)=\underset{c_{ij}^*(t)}{\arg\min}\left\{\begin{matrix}\left[\partial P/\partial\mathbf{C}(t)\right]_{ij}c_{ij}^*(t)+\\\left[\alpha\left|c_{ij}^*(t)\right|+\beta\left(c_{ij}^*(t)\right)^2\right]\end{matrix}\right\} \tag{15}$$

where $\left[\partial P/\partial\mathbf{C}(t)\right]_{ij}$ is the element in the *i*-th row and *j*-th column of matrix $\partial P/\partial\mathbf{C}(t)$.

Notably, the transformed optimization objective shown in Eq. (15) does not any contain interaction terms of elements in $\mathbf{C}(t)$. Thus, a complex multivariate optimization problem is simplified to the univariate level. According to the optimization theory for non-smooth problems [33], we can derive the final update strategy of $\mathbf{C}(t)$ as follows:

$$c_{ij}(t+1)=\begin{cases}0 & \left|\dfrac{\partial P}{\partial c_{ij}(t)}\right|\leq\alpha\\-\dfrac{1}{2\beta}\left(\dfrac{\partial P}{\partial c_{ij}(t)}-\lambda\right) & \dfrac{\partial P}{\partial c_{ij}(t)}>\alpha\\-\dfrac{1}{2\beta}\left(\dfrac{\partial P}{\partial c_{ij}(t)}+\lambda\right) & \dfrac{\partial P}{\partial c_{ij}(t)}<-\alpha\end{cases} \tag{16}$$

where the gradient can be calculated as:

$$\frac{\partial P}{\partial c_{ij}(t)}=2\left(\hat{x}_i(t)-x_i(t)\right)\left[\mathbf{X}_{[t-T:t-1]}\mathbf{W}(t)\right]_{ij} \tag{17}$$

Due to space limitations, the detailed derivation of the above optimization problem is not presented here. So far, the updating strategies of $\mathbf{W}(t)$ and $\mathbf{C}(t)$ are given in Eqs. (11) and (16). At each sample point, we can adaptively update the model parameters to accommodate the nonstationarity. In addition, the sparsity of the parameters can be controlled by adjusting *α*. Specifically, the larger the value of *α*, the sparser the $\mathbf{C}(t)$ matrix will be.

B. Overall causal metric

Although the predictive relationships between process variables can be time-varying due to nonstationarity, the overall causalities should reflect the time-independent fault mechanism. Therefore, a metric is desired to induce the overall causal relations from time-varying predictive relations. Here, we draw on the idea of parameter reliability [34] in statistical learning to measure the stability of the predictive relationship in terms of mean and standard deviation, thereby characterizing the causal relationship. On the one hand, a predictive relationship driven by a reliable causality should have a relatively high average strength, which can be represented by the average of the coefficients in the $\mathbf{C}(t)$ matrix. On the other hand, a predictive relationship that results from a significant causal dependency may slowly change over time due to nonstationarity but should not undulate wildly in the short term. This can be represented by the standard deviation of the coefficients in the $\mathbf{C}(t)$ matrix.

Considering the above two points, the following causal metric is constructed:

$$\tilde{c}_{ij} = \frac{1}{K-T}\sum_{t=T+1}^{K}\frac{\text{mean}\left\{\left|c_{ij}(t)\right|,\ldots,\left|c_{ij}(t+T-1)\right|\right\}}{\text{std}\left\{\left|c_{ij}(t)\right|,\ldots,\left|c_{ij}(t+T-1)\right|\right\}+\varepsilon} \tag{18}$$

where K denotes the number of samples; $\text{mean}\{\cdot\}$ and $\text{std}\{\cdot\}$ are the mean and standard deviation functions; ε is a constant used to avoid infinite values.

It can be seen that if the predictive relationship between two variables has higher average strength and less variation, then it will have a relatively higher metric to indicate a strong causal relation.

To provide intuitive RCD results, the root cause score needs to be quantified for each variable. A variable with a higher score is more likely to be the root cause variable for the fault. According to previous studies [35], [36] if the causal relations between a certain variable and other variables are more significantly changed after the fault, then this variable is more inclined to be a root cause variable. Accordingly, we design the following root cause scoring mechanism, for each variable $\mathbf{x}_i$, its root cause score is quantified as:

$$s(\mathbf{x}_i) = \sum_{j=1}^{N}\left|\tilde{c}_{ji}^{(n)} - \tilde{c}_{ji}^{(f)}\right| \tag{19}$$

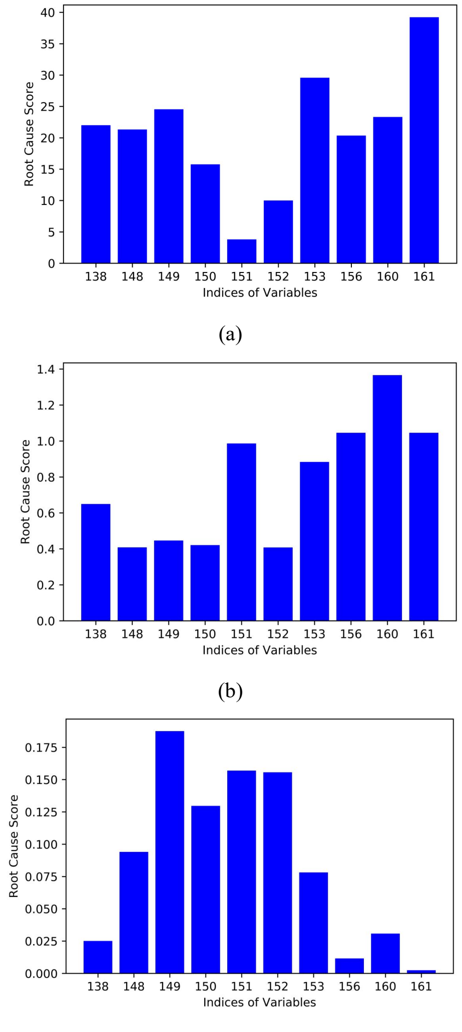
where $\tilde{c}_{ji}^{(n)}$ and $\tilde{c}_{ji}^{(f)}$ are the causal metrics from $\mathbf{x}_j$ to $\mathbf{x}_i$ extracted from the data before and after the fault, respectively.

IV. Experiments and Discussions

This section presents a real industrial example to verify the performance of the proposed method in the actual industrial application. We collected the operation data from the condensing system of a thermal power plant in Zhejiang Province. A detailed description of the thermal power plant and the condensing system can be found in [31].

There are 2820 samples and 162 variables in the adopted dataset. A blockage-induced surge in the outlet pressure of the circulating water pump B started from the 500th sample, resulting in lots of downstream variables showing abnormal trends. Specifically, there were two root cause variables, x_{156} and x_{161}, which were the two outlet pressure measurement points of the circulating pumps. Two hundred normal samples before the fault and 50 faulty samples after the fault were used for model training. Two classical RCD methods, STE [25] and LG [10], are used to provide comparisons. The PCA-based contribution plot method [5] was used to isolate the key faulty variables. The isolated variables were: $\{x_{138}, x_{148}, x_{149}, x_{150}, x_{151}, x_{152}, x_{153}, x_{156}, x_{160}, x_{161}\}$.

We visualized the diagnostic results of each method in Fig. 1. It can be seen that the proposed method correctly identifies one of the two root dependent variables, namely x_{161}. However, both STE and LG gave the wrong root cause variables. STE determined x_{160} (the outlet pressure measurement point of circulating water pump A) as the root cause variable. Although this was relatively close to the true root causes, it was still not the correct result. There were strong coupling relationships between the faulty variables in this case, so STE might be misled by this coupling relationship due to the loss of key information from the symbolic operation. Unlike STE, the RCD results given by LG are far from the actual situation. In this case, the fault was a step change that resulted in strong nonstationarity, which cannot be handled by the LG method because it required a stationarity assumption.



(a)

(b)

(c)

Fig. 1. RCD results of different methods for the condensing system example. (a) Proposed method. (b) STE. (c) LG.

Although the proposed model achieved the best performance among all methods, we were curious why it omitted one of the root dependent variables. Although the proposed model achieved the best performance among all methods, we were curious why it omitted one of the root dependent variables. After data analysis, we found that the physical distances of the two root cause measurement points were adjacent, which resulted in very close pressure values measured by the two. Calculating the Pearson correlation coefficient between the root cause variables x_{156} and x_{161}, one can find results as high as 0.99. Due to the sparse variable selection mechanism in the proposed method, it tended to focus on one of the two highly correlated variables and ignore the another. Nonetheless, since x_{156} and x_{161} were very close, the proposed method could still find the correct root cause, which verified its applicability in industrial scenarios.

V. Conclusion

In this study, we propose a time-varying predictive relation extraction model with sparsity constraints, which can capture significant GC in nonstationary processes. Also, the designed causal metric and root cause scoring mechanism can capture overall causalities from time-varying relations and provide a quantified representation of root cause variables. The proposed method can be successfully applied to the RCD task of a real industrial process in a thermal power plant. Compared with the classical RCD models, STE and LG, the proposed method correctly identified a root cause variable of the fault, verifying its diagnostic accuracy. Also, the experimental results verified its practicality in actual industrial scenes.

Acknowledgment

This work was supported in part by the National Key R&D Program of China (2021YFC2902701), the National Natural Science Foundation of China (No. 62125306), the Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (ICT2021A15), and State Key Laboratory of Synthetical Automation for Process Industries (2020-KF-21-07).

References

[1] X. Xu, J. Ding, Q. Liu and T. Chai, "A Novel Multimanifold Joint Projections Model for Multimode Process Monitoring," *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 5961-5970, 2021.

[2] W. Yu and C. Zhao, "Recursive Exponential Slow Feature Analysis for Fine-Scale Adaptive Processes Monitoring With Comprehensive Operation Status Identification," *IEEE Trans. Ind. Informat.*, vol. 15, no. 6, pp. 3311-3323, 2019.

[3] C. Zhao and H. Sun, "Dynamic distributed monitoring strategy for large-scale nonstationary processes subject to frequently varying conditions under closed-loop control," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4749-4758, 2019..

[4] C. Zhao and B. Huang, "A full-condition monitoring method for nonstationary dynamic chemical processes with cointegration and slow feature analysis," *AIChE J.*, vol. 64, no. 5, pp. 1662-1681, 2018.

[5] J. Westerhuis, S. Gurden and A. Smilde, "Generalized contribution plots in multivariate statistical process monitoring," *Chemom. Intell. Lab. Syst.*, vol. 51, no. 1, pp. 95-114, 2000.

[6] W. Yu and C. Zhao, "Sparse exponential discriminant analysis and its application to fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5931-5940, 2018.

[7] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econ., J. Econ. Soc.*, vol. 37, no. 3, pp. 424–438, 1969.

[8] H. Chen et al., "Systematic procedure for Granger-causality-based root cause diagnosis of chemical process faults," *Ind. Eng. Chem. Res.*, vol. 57, no. 29, pp. 9500-9512, 2018.

[9] C. K. I. Williams and C. E. Rasmussen, "Gaussian Processes for Regression," Advances in Neural Information Processing Systems 8, pp. 514-520, 1996.

[10] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in SIGKDD. ACM, pp. 66–75, 2007.

[11] L. A. Baccala and K. Sameshima, "Partial directed coherence: A new concept in neural structure determination," *Biol. Cybern.*, vol. 84, no. 6, pp. 463–474, 2001.

[12] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461-464, 2000.

[13] P. Duan, F. Yang, T. Chen and S. Shah, "Direct causality detection via the transfer entropy approach," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 6, pp. 2052-2066, 2013.

[14] B. Lindner, A. Lidia, B. Margret, "A systematic workflow for oscillation diagnosis using transfer entropy," *IEEE Trans. Contr. Sys. Tech.*, vol. 28, no. 3, pp. 908-919, 2019.

[15] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Phys. Rev. Lett.*, vol. 103, no. 23, pp. 238701, 2009.

[16] K. P. Murphy, "Dynamic Bayesian Networks: Representation, inference and learning," Ph.D. dissertation, *Dept. Comput. Sci., Univ. California*, Berkeley, Berkeley, CA, USA, 2002.

[17] B. Cai, L. Huang and M. Xie, "Bayesian networks in fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2227-2240, 2017.

[18] M. Amin, F. Khan and S. Imtiaz, "Fault detection and pathway analysis using a dynamic Bayesian network," *Chem. Eng. Sci.*, vol. 195, pp. 777-790, 2019.

[19] J. Yu and M. M. Rashid, "A novel dynamic Bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis," *AIChE J.*, vol. 59, no. 7, pp. 2348–2365, 2013

[20] A. Papana et al., "Identifying causal relationships in case of non-stationary time series," Department of Economics of the University of Macedonia, Thessaloniki, 2014.

[21] E. Möller et al., "Fitting of one ARMA model to multiple trials increases the time resolution of instantaneous coherence," *Biol. Cybern.*, vol. 89, no. 4, pp. 303–312, 2003.

[22] E. Möller et al., "Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models," *J. Neurosci. Methods*, vol. 105, pp. 143–158, 2001.

[23] R. Raveendran, B. Huang and W. Mitchell, "A Variational Bayesian Causal Analysis Approach for Time-Varying Systems," *IEEE Trans. Control Syst. Technol.*, vol. 29, no. 3, pp. 1191-1202, 2021.

[24] T. Schäck, M. Muma, M. Feng, C. Guan and A. M. Zoubir, "Robust nonlinear causality analysis of nonstationary multivariate physiological time series," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1213-1225, 2017.

[25] M. Staniek and K. Lehnertz, "Symbolic transfer entropy," *Phys. Rev. Lett.*, vol. 100, no. 15, pp. 158101, 2008.

[26] D. Kugiumtzis, "Partial transfer entropy on rank vectors," *Eur. Phys. J. Spec. Top.*, vol. 222, no. 2, pp. 401–420, 2003.

[27] S. Duan, C. Zhao and M. Wu, "Multiscale partial symbolic transfer entropy for time-delay root cause diagnosis in nonstationary industrial processes," *IEEE Trans. Ind. Electron.*, to be published. doi: 10.1109/TIE.2022.3161761.

[28] Y. He, Y. She and D. Wu, "Stationary-sparse causality network learning," *J. Mach., Learn. Res.*, vol. 14, pp. 3073-3104, 2013.

[29] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. R. Stat. Soc. B*, vol. 58 , pp. 267–288, 1996.

[30] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B* (*Stat. Methodol.*), vol. 67, no. 2, pp. 301–320, 2005.

[31] J. Duchi and Y. Singer, "Efficient online and batch learning using forward backward splitting," *J. Mach., Learn. Res.*, vol. 10, pp. 2899-2934, 2009.

[32] P. Song and C. Zhao, "Sparse Adjacency Forecasting and Its Application to Efficient Root Cause Diagnosis of Process Faults," *IFAC-PapersOnLine*, vol. 54, no. 3, pp. 439-444, 2021.

[33] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," Advances in Neural Information Processing Systems, 2009.

[34] V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste and C. Sterna, "Elimination of uninformative variables for multivariate calibration," *Anal. Chem.*, vol. 68, no. 21, pp. 3851-3858, 1996.

[35] W. Cheng, K. Zhang, H. Chen, G. Jiang, Z. Chen and W. Wang, "Ranking causal anomalies via temporal and dynamical analysis on vanishing correlations," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 805-814.

[36] J. Ni, W. Cheng, K. Zhang, D. Song, T. Yan, H. Chen and X. Zhang, "Ranking causal anomalies by modeling local propagations on networked systems," , in IEEE International Conference on Data Mining (ICDM), 2017, pp. 1003-1008.