On correlation analysis of bivariate alarm signals

Zijiang Yang and Jiandong Wang.

College of Engineering
Peking University
Beijing, 100871, China
{yangzj jiandong}@pku.edu.cn

Tongwen Chen

Dept. of Electrical & Computer Engineering
University of Alberta
Edmonton, Alberta, T6G 2V4, Canada
tchen@ece.ualberta.ca

Abstract—This paper studies the correlation analysis for bivariate alarm signals in order to indicate whether two alarm signals are correlated, and their cause-effect direction if they are correlated. First, seven similarity coefficients are selected as the similarity measurements for binary data, in consideration of specialties of alarm signals. The upper bounds of these similarity coefficients are theoretically established for two independent alarm signals. Second, the distribution of a so-called correlation delay is shown to be indispensable and effective for the correlation analysis. Finally, a novel correlation analysis method for alarm signals is proposed based on the correlation ratio coefficient and distribution of the correlation delay. An industrial case study is provided to illustrate the proposed correlation analysis method.

Index Terms—Alarm signals, correlation analysis, similarity coefficients, cause-effect relationship.

I. INTRODUCTION

Systems are of paramount importance to safety and efficient operation of modern industrial processes [1][10]. A well designed and efficient alarm system should follow some industrial criteria such as those in the guide from the Engineering Equipment and Materials Users’ Association (EEMUA) [6]. However, according to recent industrial surveys, operators of industrial plants often receive much more alarms than they can handle promptly. Many alarms belong to the so-called consequential alarms that are triggered by other root-cause alarms. Due to the increasing complexity and varying dynamics in the modern industrial processes, usually it is a difficult and time-consuming task for operators to classify the consequential and root-cause alarms solely based on process knowledge and operation experience. Hence, it would be valuable to have a systematic approach to perform the classification in an automatic manner, and design advance alarm management systems to improve the efficiency of dealing with these alarms. The correlation analysis is one of the most fundamental statistical tools to reach the above objectives.

The correlation analysis for alarm signals has received increasing attentions lately. Dahlstrand [5] used multilevel flow models based on the process knowledge to perform consequence analysis to find the root cause of alarms. Brooks et al. [2] exploited the parallel coordinate to capture the relationship of multivariate alarm signals and to obtain dynamically varying alarm limits, providing the historical data containing the information of best operating zones. Yang et al. [11] analyzed the discrepancy of correlation between process data and alarm data and used this information to optimize the alarm limits. Kondaveeti et al. [8] provided an alarm similarity color map to group similar alarms together based on the Jaccard similarity coefficient after padding alarm sequence with extra 1’s to existing alarm occurrences. Noda et al. [9] conducted correlation analysis between alarms and operation events to identify sequential alarms and unnecessary operations. Yang et al. [12] generated pseudo continuous time series from the original binary alarm data and used a correlation color map of the pseudo data to show the cluster of correlated variables.

The objective of this paper is to perform the correlation analysis for alarm signals in order to indicate whether two alarm signals are correlated or not, and their cause-effect direction if they are correlated. The contribution is three-fold: First, a similarity measurement, namely, the correlation ratio coefficient, is selected among twenty-two existing similarity coefficients for binary data. The selection is on the basis of some desired properties arising from specialties of alarm signals. In particular, the upper bounds of seven similarity coefficients are theoretically established for independent alarm signals. Second, a so-called correlation delay, which is the time shift for two alarm signals to achieve the maximum correlation ratio coefficient, is shown to be indispensable and effective for the correlation analysis. Third, a novel correlation analysis method for alarm signals is proposed based on the correlation ratio coefficient and distribution of the correlation delay.

The rest of the paper is organized as follows. Section II selects the correlation ratio coefficient as the similarity measure suitable for alarm signals. Section III discusses the necessity of looking at the distribution of the correlation delay, and proposes a correlation analysis method for alarm signals. Section IV provides numerical and industrial examples. Finally, some concluding remarks are given in Section V.

*This work was partially supported by Shandong Electric Power Research Institute, NSERC, and the National Natural Science Foundation of China under Grants No. 60704031 and No.61061130559.
II. SIMILARITY COEFFICIENTS FOR ALARM SIGNALS

This section summarizes twenty-two similarity coefficients for binary sequences and evaluates their suitability for alarm signals based on certain criteria. The correlation ratio coefficient is recommended as the final choice.

A. Similarity measurements for binary sequences

As alarm signals are composed by only ‘1’s and ‘0’s, the similarity measurements for binary sequences are applicable to alarm signals. There are perhaps seventy-six statistical coefficients as the similarity measurements in the literature [4]. Here twenty-two of them based on [3] (Table I therein) and [7] (Table 3 therein) are summarized in Table I. The mathematical symbols in Table I are defined as follows:

- \( N \): the data length of the two sequences,
- \( N_1 \): the number of 1s in the first sequence,
- \( N_2 \): the number of 1s in the second sequence,
- \( A \): the number of 0s appeared simultaneously in both sequences,
- \( C \): the number of 1s appeared simultaneously in both sequences,
- \( E_1 \): the number of 1s in the first sequence corresponding to 0s in the second sequence,
- \( E_2 \): the number of 1s in the second sequence corresponding to 0s in the first sequence.

Basic properties of the coefficients in Table I have already been studied, see, e.g., [3]. In particular, some coefficients focus on the matching of ‘1’ to ‘1’, while some pay more attention to the matching of other types, e.g., ‘0’ to ‘0’. Hence, we need to choose coefficients more suitable to alarm signals.

B. Two formulations of alarm signals

It is a common practice in the industry that when a process signal goes into the alarm state, the corresponding alarm signal changes the value from ‘0’ to ‘1’, when the process signal runs from the alarm state to the non-alarm one, the alarm signal makes an opposite switch from ‘1’ to ‘0’ and stays with the value of ‘0’ throughout the period of being in the non-alarm state. Thus, the similarity coefficients for alarm signals should be the ones focusing on the matching of ‘1’ to ‘1’. However, there are two possible formulations for the alarm signal when the process signal is in the alarm state. The first formulation is to let the alarm signal take the value of ‘1’ throughout the period in the alarm state, while the alarm signal in the second formulation takes the value of ‘1’ only at the time instant when the process signal goes into the alarm state from the non-alarm state. To be precise, let \( x(n) \) and \( x_{tp} \) be the process signal and its associated high-alarm trippoint value, respectively. For the first formulation, the alarm signal is

\[
x_a(n) = \begin{cases} 
1, & \text{if } x(n-1) < x_{tp} \text{ and } x(n) \geq x_{tp} \\ 
0, & \text{otherwise}
\end{cases}
\]

The alarm signal in the second formulation is

\[
x_a(n) = \begin{cases} 
1, & \text{if } x(n) \geq x_{tp} \\ 
0, & \text{otherwise}
\end{cases}
\]

Both formulations have been adopted in literature, see, e.g., [8][11][12].

The second formulation is more suitable to generate alarm signals for the subsequence correlation analysis. It provides the information on the time instant when the process signal goes into the alarm state from the non-alarm state, which is crucial to tell if two alarm sequences occur in a correlated manner. By contrast, this information is very likely to be overwhelmed for the first formulation by many other ‘1’s arisen from the period that the process signal stays in the alarm state. Since the similarity coefficients for alarm signals are based on the matching of ‘1’ to ‘1’, the first formulation will overestimate the similarity coefficients and may lead to incorrect conclusions. Therefore, the second formulation is adopted in the sequel.

C. Selection of similarity coefficients

With alarm signals generated as that in (1), we are ready to select a similarity coefficient suitable for alarm signals among

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>( \frac{N_1 \cap N_2}{N_1 + N_2 - N_1 \cap N_2} )</td>
</tr>
<tr>
<td>Simple M.</td>
<td>( \frac{C}{N_1 + N_2 - A + C} )</td>
</tr>
<tr>
<td>Dice</td>
<td>( \frac{N_1 \cap N_2}{N_1 + N_2} )</td>
</tr>
<tr>
<td>1st Kulcz.</td>
<td>( \frac{N_1 \cap N_2 - 2(C + A)}{N_1 \cap N_2 + C + A} )</td>
</tr>
<tr>
<td>2nd Kulcz.</td>
<td>( \frac{N_1 \cap N_2}{N_1 \cap N_2 + C + A} )</td>
</tr>
<tr>
<td>Otsuka</td>
<td>( \frac{1}{\sqrt{N_1 N_2}} )</td>
</tr>
<tr>
<td>Correlation ratio</td>
<td>( \frac{C^2}{N_1 N_2} )</td>
</tr>
<tr>
<td>Simpson</td>
<td>( \frac{A}{N_1 N_2} )</td>
</tr>
<tr>
<td>Braun-Bl</td>
<td>( \frac{C}{N_1 N_2} )</td>
</tr>
<tr>
<td>Rogers &amp; Tan.</td>
<td>( \frac{E_1 E_2}{(C + D)(N_1 \cap N_2)} )</td>
</tr>
<tr>
<td>Hamann</td>
<td>( \frac{E_1 E_2}{(C + D)(N_1 \cap N_2)} )</td>
</tr>
<tr>
<td>Yale</td>
<td>( \frac{E_1 E_2}{(N_1 \cap N_2)(C + A)} )</td>
</tr>
<tr>
<td>Phi</td>
<td>( \frac{E_1 E_2}{(N_1 \cap N_2)(C + A)} )</td>
</tr>
<tr>
<td>Resem. Eq.</td>
<td>( 1 - z, \text{ where } z \text{ is defined in } \frac{1}{N_1 N_2} \sqrt{C + A} )</td>
</tr>
<tr>
<td>Coeff. Diff.</td>
<td>( 1 - \frac{N_1 N_2}{N_1 N_2 + 2N_1 N_2 - C - A} )</td>
</tr>
<tr>
<td>No. Feat. Diff.</td>
<td>( C + A )</td>
</tr>
<tr>
<td>Sokal Dist (a)</td>
<td>( \frac{E_1 + E_2}{N_1 + N_2 - C - A} )</td>
</tr>
<tr>
<td>Sokal Dist (b)</td>
<td>( 1 - \frac{N_1 N_2}{N_1 + N_2 - C - A} )</td>
</tr>
<tr>
<td>Raw</td>
<td>( \frac{C + A}{N_1 N_2} )</td>
</tr>
<tr>
<td>Baroni Urbani</td>
<td>( \frac{C + A}{N_1 N_2} )</td>
</tr>
<tr>
<td>Kulzinski</td>
<td>( \frac{C + A}{N_1 N_2} )</td>
</tr>
</tbody>
</table>
those in Table I. Because alarm signals may be correlated in a dynamic manner, the similarity coefficients in Table I are
generalized as
\[ S_{\text{name}, x_a, y_a} = \max_{\tau} \rho_{\text{name}, x_a, y_a}(\tau). \tag{2} \]

where the subscript name stands for the name of any similarity coefficient in Table I, and \( \rho_{\text{name}, x_a, y_a}(\tau) \) is the value of a similarity coefficient between an alarm signal \( x_a(n) \) and another one \( y_a(n + \tau) \). Here \( y_a(n + \tau) \) is obtained by shifting the alarm signal \( y_a(n) \) forward by \( \tau \) samples if \( \tau \geq 0 \) and backward by \( |\tau| \) samples if \( \tau < 0 \). For finite data lengths, a zero-padding strategy is exploited to make \( x_a(n) \) and \( y_a(n + \tau) \) having the same number of samples.

By considering the characteristic of alarm signals, it would be desirable for a selected similarity coefficient to have the following properties:

(a) it focuses on the matching of ‘1’ to ‘1’ only;
(b) it takes a value inside the interval \([0, 1]\);
(c) its upper bound should be as small as possible for the case that two alarm signals are independent.

Property (a) is a natural choice for the alarm signals generated as that in (1) where the matching of ‘1’ to ‘1’ implies that alarms are tripped on in a synchronized manner. Property (b) is a standard requirement so that the similarity coefficient can be easily interpreted: If the similarity coefficient takes a value close to 1 (0), then two alarm signals are strongly (weakly) correlated. The upper bound of the similarity coefficient in Property (c) is useful to determine a threshold for a hypothesis test on whether two alarm signals are correlated or not. Since the similarity coefficient closer to 0 means a weaker correlation, the upper bound should be as small as possible. Otherwise, if the upper bound is quite large, e.g., 1/2 for Simpson coefficient in Table II, then it would be against intuition to interpret such a high value standing for two alarm signals as being uncorrelated.

Based on the three properties, we can now select a similarity coefficient suitable for alarm signals among those in Table I. Property (a) rules out the similarity coefficients containing the term \( A \), because alarm signals contain much more ‘0’s than ‘1’s and the term \( A \) will play a dominant role to overwhelm the matching of ‘1’ to ‘1’. The value ranges of similarity coefficients in Table I have already been investigated in [3] so that those similarity coefficients do not satisfy Property (b) can be readily discarded. In summary, the similarity coefficients satisfying Properties (a) and (b) include Jaccard, Dice, 2nd Kulcz, Otsuka, correlation ratio, Simpson and Braun-Bl.

For Property (c), we establish the upper bounds of the above seven similarity coefficients for two independent alarm signals; the upper bounds are given in Table II. The correlation ratio has the smallest upper bound 1/16, much smaller than others.

Table II

<table>
<thead>
<tr>
<th>Similarity coefficient</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>5/7</td>
</tr>
<tr>
<td>Dice</td>
<td>5/7</td>
</tr>
<tr>
<td>2nd Kulcz</td>
<td>5/7</td>
</tr>
<tr>
<td>Otsuka</td>
<td>5/7</td>
</tr>
<tr>
<td>Simpson</td>
<td>5/7</td>
</tr>
<tr>
<td>Braun-Bl</td>
<td>5/7</td>
</tr>
</tbody>
</table>

**Proposition 1.** If two alarm signals \( x_a(n) \) and \( y_a(n) \) are independent to each other and the corresponding process signals \( x(n) \) and \( y(n) \) are independent and identically distributed (IID), then the upper bounds of correlation ratio, Jaccard, Dice, 2nd Kulcz, Otsuka, Simpson and Braun-Bl are those in Table II.

**Proof of Proposition 1.** Here we prove the upper bound of the correlation ratio coefficient, namely, \( \rho_{\text{corr}, x_a, y_a} \leq \frac{1}{16} \). For Jaccard, Dice, 2nd Kulcz, Otsuka, Simpson and Braun-Bl coefficients, their upper bounds can also be theoretically established in a similar manner; the proofs are omitted due to space limitation. Consider the collected samples of alarm signals \( \{x_a(n), y_a(n)\}_{n=1}^N \) and their alarm trippoint values \( x_{tp} \) and \( y_{tp} \). Define \( P_1 \) and \( P_2 \) as the probabilities of \( x(n) \) and \( y(n) \) greater than their own alarm trippoint values \( x_{tp} \) and \( y_{tp} \), respectively, i.e., \( P_1 = P(x \geq x_{tp}) \) and \( P_2 = P(y \geq y_{tp}) \). Since \( x(n) \) and \( y(n) \) are IID, and \( x_a(n) \) and \( y_a(n) \) are assumed to be independent to each other, (1) implies that \( N_1, N_2 \) and \( C \), defined in Section II-A, are approximately equal to \( NP_1(1 - P_1), NP_2(1 - P_2) \) and \( NP_1(1 - P_1)NP_2(1 - P_2) \), respectively. Thus, the correlation ratio coefficient becomes

\[
\rho_{\text{corr}, x_a, y_a} = \frac{C^2}{N_1N_2} \approx \frac{[NP_1(1 - P_1)NP_2(1 - P_2)]^2}{NP_1(1 - P_1)NP_2(1 - P_2)} = P_1(1 - P_1)P_2(1 - P_2). 
\]

Using the triangular inequality, we have

\[
\rho_{\text{corr}, x_a, y_a} \leq \left( \frac{P_1 + 1 - P_1}{2} \right)^2 \left( \frac{P_2 + 1 - P_2}{2} \right)^2 = \frac{1}{16}. 
\]

The equality holds if and only if \( P_1 = 1 - P_1 \) and \( P_2 = 1 - P_2 \), leading to \( P_1 = P_2 = 0.5 \).

Remark #1: The correlation ratio coefficient is used for the subsequent correlation analysis, because it has the smallest upper bound. The Jaccard coefficient as well as other similarity coefficients in Table II are equally applicable, too.
III. CORRELATION ANALYSIS

This section discusses the necessity of the distribution of the correlation delay, and proposes a novel method to perform correlation analysis on alarm signals. The method can indicate whether two alarm signals are statistically correlated or not, quantify their correlation level if they are correlated, and give their cause-effect relationship.

A. Correlation delay

The so-called correlation delay, denoted as $\tau_s$, is defined as the value of delay $\tau$ achieving the maximum value of the correlation ratio coefficient sequence $\rho_{\text{corr},x_a,y_a}(\tau)$. That is, (2) can be written here as

$$S_{\text{corr},x_a,y_a} = \rho_{\text{corr},x_a,y_a}(\tau_s).$$  (3)

The correlation delay $\tau_s$ is indispensable for correlation analysis for two reasons. First, the similarity coefficient can only tell whether two alarm signals are correlated and their correlation level, i.e., based on the upper bound of the similarity coefficient $S_{\text{corr},x_a,y_a}$ in Proposition 1, it is concluded that $x_a(n)$ and $y_a(n)$ are correlated if $S_{\text{corr},x_a,y_a}$ is greater than the upper bound $\frac{1}{16}$. The physical meaning of $S_{\text{corr},x_a,y_a}$ being larger than $\frac{1}{16}$ is the statistical level that $x_a(n)$ and $y_a(n + \tau_s)$ switch from ‘0’ to ‘1’ simultaneously. However, $S_{\text{corr},x_a,y_a}$ cannot give the cause-effect relationship. Thus, it has to be complemented with the information of correlation delay $\tau_s$. That is, if $\tau_s > 0$ then $x_a(n)$ affects $y_a(n)$ with the delay $\tau_s$, and vice versa. The second reason is less obvious. If $S_{\text{corr},x_a,y_a} < \frac{1}{16}$, then it does not necessarily mean that $x_a(n)$ and $y_a(n)$ are uncorrelated. This is due to the randomness induced by noise and the fact that the upper bound is established based on certain assumptions in Proposition 1, while it is very possible that $S_{\text{corr},x_a,y_a}$ for two correlated alarm signals $x_a(n)$ and $y_a(n)$ is less than $\frac{1}{16}$. In this case, we need to look at the distribution of correlation delay $\tau_s$ to check whether the distribution is concentrated to a small interval. The necessity of looking at the distribution of $\tau_s$ is illustrated in the following numerical examples.

Remark #2: It is important to detect the correlation relationship, even if the current value of $S_{\text{corr},x_a,y_a}$ is small, and $x_a(n)$ and $y_a(n)$ look like to be weakly correlated. Such a small value may be caused by improper setting of alarm trip point values, and the corresponding process signals $x(n)$ and $y(n)$ are actually strongly correlated. By realizing the presence of correlation, we may re-design the alarm trip point values to greatly improve the performance of alarm systems.

Example 3a. The process signal $y(n)$ is generated as $y(n) = x^2(n - 5) + e(n)$, where $x(n)$ is white noise with standard Gaussian distribution, and $e(n)$ is another Gaussian white noise with mean value 0 and standard deviation 0.05. Here $x(n)$ and $e(n)$ are independent to each other. The trip point values of $x(n)$ and $y(n)$ are $x_{tp} = 1.8$ and $y_{tp} = 0.5$, respectively. Fig. 1-(a), (b) and (c) respectively present the similarity coefficient $S_{\text{corr},x_a,y_a}$, the correlation ratio coefficient sequence $\rho_{\text{corr},x_a,y_a}(\tau)$ for $\tau \in [-30, 30]$, and the distribution of $\tau_s$, obtained in 100 Monte Carlo simulations with the data length $N = 10000$. The sample average of $S_{\text{corr},x_a,y_a}$ is 0.04 and the correlation delay $\tau_s$ is concentrated to the right time delay 5. As stated earlier in Remark #2, the small value of $S_{\text{corr},x_a,y_a}$ may be caused by improper setting of alarm trip point values. If $x_{tp} = 2.0$ and $y_{tp} = 5.0$, then the sample average of $S_{\text{corr},x_a,y_a}$ is increased to 0.27 in another set of 100 Monte Carlo simulations.

Example 3b. The process signals $x(n)$ and $y(n)$ are independent to each other, both taking standard Gaussian distribution. The trip point values are $x_{tp} = 0.9$ and $y_{tp} = 0.5$. Fig. 2 is the counterpart of Fig. 1, based on 100 Monte Carlo simulations with the data length $n = 10000$. The sample average of $S_{\text{corr},x_a,y_a}$ is 0.04 and $\tau_s$ is not concentrated to a small interval; hence, the distribution of $\tau_s$ correctly indicates that $x_a(n)$ and $y_a(n)$ are uncorrelated.

In Examples 3a and 3b, the values of similarity coefficient $S_{\text{corr},x_a,y_a}$ are almost the same, less than the upper bound $\frac{1}{16}$, but the actual relation between $x_a(n)$ and $y_a(n)$ is completely different. The difference can be effectively found via the distribution of $\tau_s$. 

533
B. Number of alarms

In order to reach reliable estimates of the similarity coefficient $S_{\text{corr}, x_a, y_a}$ and the correlation delay $\tau_s$, the numbers of alarms in $x_a(n)$ and $y_a(n)$ cannot be too small; otherwise, the alarms tripped by noise could easily deteriorate the reliability of the estimates. A natural question is: How many ‘1’s in $x_a(n)$ and $y_a(n)$ are required?

We propose a rule of thumb as the answer: The numbers of ‘1’s in the alarm signals $x_a(n)$ and $y_a(n)$, generated via (1), should be at least 30. This rule of thumb is obtained based on a numerical study as follows. Suppose that $y_a(n) = x_a(n) + e_a(n)$, where both $x_a(n)$ and $e_a(n)$ are binary alarm signals. Let the number of ‘1’s in $y_a(n)$ be a positive integer $K$. The position of each ‘1’ in $y_a(n)$ is uniformly selected and is denoted as $n_a$. Then, ‘1’ appears either at $x_a(n_a)$ with probability $p$ or at $e_a(n_a)$ with probability $1 - p$ for a real number $p \in [0, 1]$, i.e.,

$$\Pr(x_a(n_a) = 1, e_a(n_a) = 0) = p,$$
$$\Pr(x_a(n_a) = 0, e_a(n_a) = 1) = 1 - p.$$ 

Therefore, the true similarity coefficient is approximately equal to

$$S_{\text{corr}, x_a, y_a} \approx \frac{(Kp)^2}{Kp} = p.$$ 

The approximation is due to the possible rounding error of $Kp$ to the integer number of alarms in $x_a(n)$. It is expected that the estimated similarity coefficient may be quite far away from the true value $p$ due to the noise effects for small values of $K$, and get closer to the true value while $K$ increases. The expectation is numerically validated in Fig. 3. 100 Monte Carlo simulations are performed for each value of $K$, with a fixed data length $N = 10000$ of $x_a(n)$ and $y_a(n)$. Based on Fig. 3, we choose $K = 30$ as a threshold of trusty, that is, the estimated similarity coefficient is expected to be reliable when there are no less than 30 ‘1’s in $x_a(n)$ and $y_a(n)$.

C. Correlation analysis method

This subsection proposes a novel method to perform correlation analysis on alarm signals, based on the correlation ratio coefficient sequence $\rho_{\text{corr}, x_a, y_a}(\tau)$, the similarity coefficient $S_{\text{corr}, x_a, y_a}$ and the distribution of the correlation delay $\tau_s$.

The proposed method consists of the following steps:

1) Generate the alarm signals $x_a(n)$ and $y_a(n)$ as described in (1) from the process signals $x(n)$ and $y(n)$ with the associated alarm tripping values $x_{tp}$ and $y_{tp}$, respectively.

2) Separate the collected data of $x_a(n)$ and $y_a(n)$ into several data segments. Each segment is determined to make the smaller number of alarms in $x_a(n)$ and $y_a(n)$ equal to 30 (the rule of thumb in Section III-B). Denote the $i$-th separated data segment as $D_i$.

3) Calculate the correlation ratio coefficient sequence $\rho_{\text{corr}, x_a, y_a}(\tau)$ for $x_a(n)$ and $y_a(n)$ with $\tau \in [-L, L]$ for each separated data segment, and obtain the similarity coefficient $S_{\text{corr}, x_a, y_a}$ and correlation delay $\tau_s$ in (2), denoted as $S_{\text{corr}, x_a, y_a}(D_i)$ and $\tau_s(D_i)$, respectively. Here $L$ is a user-selected positive integer.

4) Investigate whether the distribution of $\tau_s$ is concentrat- ed to a small interval. The concentration check is based on the probability of $\tau_s$ lying in a small range around the mode of $\tau_s(D_i)$: If the inequality

$$\Pr(|\tau_s - \text{mode}(\tau_s(D_i))| \leq \tau_0) \geq \gamma$$

holds, then we conclude that the distribution of $\tau_s$ is concentrated. Here $\tau_0$ is a small positive integer and $\gamma$ is a positive real number less than 1, e.g., $\tau_0 = 5$ and $\gamma = 0.9$. The mode value of $\tau_s$, denoted as $\tau_s^0 := \text{mode}(\tau_s(D_i))$, is taken as the estimated correlation delay between $x_a(n)$ and $y_a(n)$, and the mean value of $S_{\text{corr}, x_a, y_a}(D_i)$, denoted as $S_{\text{corr}, x_a, y_a}^0$, is taken as the estimated similarity coefficient.

Once the above four steps have been implemented, we can conclude if two alarm signals are statistically correlated, quantify their correlation level if they are correlated, and give cause-effect causality relationship. That is, if the distribution of $\tau_s$ is concentrated to a small interval, then $x_a(n)$ and $y_a(n)$ are correlated at a level measured by the similarity coefficient $S_{\text{corr}, x_a, y_a}^0$; in addition, the correlation delay $\tau_s^0$ tells the cause-effect relation between $x_a(n)$ and $y_a(n)$: If $\tau_s^0 > 0$, then $x_a(n)$ affects $y_a(n)$; if $\tau_s^0 < 0$, then $y_a(n)$ affects $x_a(n)$.

IV. INDUSTRIAL CASE STUDY

This section provides an industrial case study to illustrate the effectiveness of the proposed correlation analysis method. Real-time measurements of two process variables $\{x(n), y(n)\}_{n=1}^N$ in a large-scale petro-chemical plant in Jiangsu Province, China, are collected with the sampling period of 1 minute. The data length is $N = 150470$, standing for the operating period about 104 days. The alarm trippoint
values are \( x_{tp} = 0.55 \) and \( y_{tp} = 0.55 \). Some part of the collected data \( \{x(n), y(n)\}_{n=1}^{N} \) with \( x_{tp} \) and \( y_{tp} \) are presented in Fig. 4. The proposed correlation analysis method in Section III is implemented as follows: Step 1 generates alarm signals \( \{x_{a}(n), y_{a}(n)\}_{n=1}^{N} \) via (1). In Step 2, the collected data are separated into 25 data segments, where the smaller number of alarms in \( x_{a}(n) \) and \( y_{a}(n) \) in each segment is equal to 30, as shown in Fig. 5-(d). Step 3 calculates the correlation ratio coefficient sequence \( \rho_{corr,x_{a},y_{a}}(\tau) \) for \( \tau \in [-20,20] \), as given in Fig. 5-(a), and the similarity coefficient \( S_{corr,x_{a},y_{a}}(D_{i}) \) and correlation delay \( \tau_{s}(D_{i}) \) in (3) for each data segment \( D_{i} \), as given in Fig. 5-(b). There are some minor variations among these \( S_{corr,x_{a},y_{a}}(D_{i}) \); their average is \( S_{corr,x_{a},y_{a}}^{0} = 0.1697 \). In Step 4, the distribution of the correlation delay \( \tau_{s}(D_{i}) \) is presented in Fig. 5-(c). Clearly, the distribution is concentrated around the mode value \( \tau_{s}^{0} = 0 \). Hence, it is concluded that \( x_{a}(n) \) and \( y_{a}(n) \) are correlated at the level measured by \( S_{corr,x_{a},y_{a}}^{0} = 0.1697 \). The cause-effect relationship between \( x_{a}(n) \) and \( y_{a}(n) \) cannot be decided, as \( \tau_{s}^{0} = 0 \). The conclusion is consistent with the available process knowledge on \( x(n) \) and \( y(n) \); they are the measurements of fuel gas pressures in two connected pipes. Thus, the alarm signals \( x_{a}(n) \) and \( y_{a}(n) \) are expected to be correlated in some manner; in addition, the pressure variation changes very quickly so that it is rather difficult to tell the cause-effect relationship based on measurements with the relatively slow sampling rate of 1 sample per minute.

V. CONCLUSION

This paper studied the problem of correlation analysis for alarm signals. Seven similarity coefficients in the literature were selected based on the properties desired for alarm signals. The upper bounds of these similarity coefficients are theoretically established in Proposition I for two independent alarm signals. The similarity coefficient \( S_{corr,x_{a},y_{a}} \) based on the correlation ratio coefficient was chosen as the one used for the subsequent correlation analysis for alarm signals. The distribution of the correlation delay \( \tau_{s} \) defined in (3) was shown to be indispensable and effective in the correlation analysis. Based on \( S_{corr,x_{a},y_{a}} \) and the distribution of \( \tau_{s} \), a novel correlation analysis method for alarm signals was proposed.

The correlation delay \( \tau_{s} \) provides valuable information of the cause-effect relationship between two alarm signals. As our future work, we need to exploit the information to classify root-cause alarms away from consequential alarms so that the number of nuisance alarms can be effectively reduced.

REFERENCES