Correlation Analysis of Alarm Data and Alarm Limit Design
for Industrial Processes

Fan Yang, Member, IEEE, Sirish L. Shah, Member, IEEE, and Deyun Xiao

Abstract—The object of alarm rationalization is to check if alarms are set correctly with respect to high and low limits and also to ensure that critical alarms are not missed and that there are as few false alarms as possible. This paper outlines a procedure based on the similarity of correlation maps of physical process variables and their alarm history in combination with process connectivity information through causal maps to suggest optimal alarm settings. The process of correlation analysis clearly takes into account the multivariate nature of the process and thereby reduces the number of false and missed alarms.

I. INTRODUCTION

MOST processes are highly interacting because of heat integration to save energy. As a result the process variables are also correlated due to causality or cause and effect relationships. Alarm systems are designed and developed to monitor the system state and also collect all the alarm data for analysis and rationalization [1]. In most modern applications, the number of alarms is very large and often beyond the operators’ ability to take actions. This is because computer control allows variables to be configured for alarm monitoring relatively easily and with many variables under monitoring, it is not easy to always acknowledge all the alarms. Alarms are also often ignored because there are often many nuisance alarms including chattering, fleeting, and stale alarms [2]. The objective in many alarm rationalization exercises is to reduce the number of alarms by improving the alarm system design. The objective in this paper is to develop tools to help the alarm rationalization process. In particular the focus of this study is to see if the correlation of the process variables is reflected in the correlation of the alarmed variables and in this way ensure that alarm limits are configured correctly to deliver minimal level of false alarms and missed alarms. An equally important objective of this study is to also develop tools that will allow identification of the root cause of alarms.

The correlation color map, in the temporal domain, is a graphical visualization of relationship between all the process variables [3]. Similarly, the corresponding binary alarm data of the same variables can also be used to generate a correlation color map. However, these two color maps are not necessarily the same because alarm data is converted from process data depending on the alarm limits. By proper configuration of the alarm limits, the correlation maps of the physical process variables and the alarm variables can reflect the multivariate nature of the process. In this way fault detection and diagnosis would be based on not singular trends of data but on multivariate trends of data.

To set alarm limits, missed alarm and false alarm are two considerations that come into play. In general, one cannot reduce both of them simultaneously by changing the limits because there is a trade-off between these two considerations [4]. Missed alarms can result in severe consequences; false alarms can also result in unneeded diversion of attention and perhaps costly or harmful intervention. There are several ways to trade off missed and false alarms. The Bayesian detector minimizes an average of the two probabilities, the minimax detector minimizes the maximum of them, and the Neyman-Pearson detector minimizes the probability of missed alarms under an upper-bound constraint on the probability of false alarms [5]. In all these methods, subjective factors take effect, so one cannot decide which one is the best. The correlation analysis provides an alternative way for choosing the limits because a good limit should differentiate two states of conditions. Thus we will analyze the discrepancy of correlation between process data and alarm data and use this to optimize the limits. The objective is to set the limits to reflect the real correlation properly so that the abnormal states can be propagated along the paths. In this way, the two probabilities can be balanced to a tolerable extent.

This paper is organized as follows. In Section II, the correlation obtained from process data and alarm data are compared theoretically and intuitively. Based on the comparison of correlation between the physical variables and the corresponding alarm limits, the alarm limits can be selected, as shown in Section III. For an industrial process, a causal map of process connectivity should be obtained and then the links can be considered one by one according to a specific order. The generation method of causal map and the choice of optimization method for selecting alarm limits for
all the variables are introduced in Section IV. A brief case study based on industrial data from Suncor Energy has been made to illustrate and validate the proposed method. Due to lack of space it is not possible to provide all details pertaining to this case study. These results will be discussed in full at the conference presentation. Conclusions and limitations of this method are summarized in the last section.

II. DIFFERENCE BETWEEN CORRELATION OF ALARM DATA AND PROCESS DATA

Two process variables can be dependent or independent, which is a reasonable concept of causality. In order to describe the dependency between two process variables, correlation is a natural criterion. However, it should be noted that dependency and correlation are not equivalent. If two random variables are independent, then they are uncorrelated. If they are correlated, then they are not independent. If there is causality between two variables, then one is dependent on the other. If they are independent, then there is no causality between them.

For example, two variables may be correlated but there may be another variable that is the common parent of them. Thus, if its value is given, then the two dependent variables can be noncausal or independent with each other despite the correlation.

Anyway, in order to know the causality and dependency, correlation is an important and efficient criterion. Hence covariance and correlation coefficient are often used to describe the relationships between two time series of process variables. Consider the process variable $x$ and $y$ with mean (or expectations) $\mu_x$ and $\mu_y$, and standard deviations $\sigma_x$ and $\sigma_y$ respectively, thus their covariance and Pearson’s correlation coefficient are:

$$\text{Cov}(x, y) = \text{E}[(x - \mu_x)(y - \mu_y)] = \text{E}[xy] - \mu_x\mu_y,$$

$$\rho_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}. \quad (1)$$

In alarm systems, the corresponding alarm data are a binary series in which “0” means no alarm and “1” means alarm. Thus intuitively one can also compute the covariance and correlation coefficient for the binary time series of alarm data. Theoretical derivation is carried out next.

Without loss of generality, the PVHI (process variable is high) alarm limit is considered first. A PVHI alarm is triggered if the value of a process variable exceeds the high alarm limit. The time series of alarm data corresponding to $x$ and $y$ are defined as $x_a$ and $y_a$. The covariance can be computed as

$$\text{Cov}(x_a, y_a) = \text{E}[x_a y_a] - \mu_{x_a}\mu_{y_a}. \quad (2)$$

Assume that the high limits of $x$ and $y$ are $x_0$ and $y_0$ respectively. By using the property of conditional expectation, we have

$$\text{Cov}(x_a, y_a) = \text{E}[x_a y_a | x > x_0, y > y_0]P(x > x_0, y > y_0) + \text{E}[x_a y_a | x \leq x_0, y > y_0]P(x \leq x_0, y > y_0) + \text{E}[x_a y_a | x > x_0, y \leq y_0]P(x > x_0, y \leq y_0) + \text{E}[x_a y_a | x \leq x_0, y \leq y_0]P(x \leq x_0, y \leq y_0) \quad (3)$$

and

$$\text{Cov}(x_a, y_a) = P(x > x_0)P(y > y_0) = 1 \cdot P(x > x_0, y > y_0) + 0 + 0 + P(x > x_0)P(y > y_0). \quad (4)$$

The variance of $x_a$ is

$$\sigma_{x_a}^2 = \text{E}[(x_a - \mu_{x_a})^2] = \text{E}[(x_a)^2] - (\mu_{x_a})^2 = P(x > x_0) - [P(x > x_0)]^2. \quad (5)$$

Let

$$P(x > x_0) = p_1, \quad P(y > y_0) = p_2, \quad (6)$$

then the variance of $x_a$ and $y_a$ is $p_1 - p_1^2$ and $p_2 - p_2^2$. The corresponding correlation coefficient of the alarm variables is given below [6].

$$\rho_{x_a y_a} = \frac{\text{Cov}(x_a, y_a)}{\sigma_{x_a} \sigma_{y_a}} = \frac{P(x > x_0, y > y_0) - p_1 p_2}{\sqrt{p_1 - p_1^2} \sqrt{p_2 - p_2^2}}. \quad (7)$$

For process data or alarm data, the correlation coefficient is between -1 and 1. For example, if $x_a$ and $y_a$ are independent ($\rho_{xy} = 0$ ), then $P(x > x_0, y > y_0)$ is equal to $P(x > x_0)P(y > y_0)$ or $p_1 p_2$, thus $\rho_{x_a y_a} = 0$. A positive correlation coefficient means a positive correlation, and a negative correlation coefficient means a negative correlation.

In general, the mapping from the correlation coefficient of two process variables to that of the corresponding alarm data is non-monotone. For example, suppose $x$ has a standard Gaussian distribution with mean 0 and variance 1. Let

$$y = \begin{cases} -x, & |x| \leq c, \\ x, & |x| > c, \end{cases} \quad (8)$$

where $c$ is a positive number. The correlation is a continuous function of $c$. The probability distribution of $y$, however, is the same as that of $x$ and does not change with respect to $c$. In this case, $x$ and $y$ can be correlated or uncorrelated according to $c$, but they are clearly not independent, since $x$ completely determines $y$. We conclude that: Assume that $p_1$ and $p_2$ are
constant. If \( P(x > x_0, y > y_0) \) is monotone with respect to the correlation coefficient \( \rho_{xy} \) between \( x \) and \( y \), then the mapping from correlation coefficient of the process data to that of the alarm data is monotone.

In most real cases, this mapping is monotone because the normal data and abnormal data can be differentiated distinctly if the limit is chosen within the valley of the probability density function as shown in Fig. 1 (dotted line). The solid lines are conditional probability density function with respect to the normal state and abnormal state. For the extreme case, the process data are similar to the alarm data which is almost “no alarm”/“alarm” binary series, then the correlation of process data and alarm data are the same.

III. ALARM LIMIT OPTIMIZATION BY CORRELATION ANALYSIS BETWEEN TWO VARIABLES

Correlation of alarm data is usually less than the correlation of corresponding process data because of the loss of quantitative values. The optimal limits can be chosen so that the correlation of alarm data is close to the real correlation of the corresponding process variables. In this way, the correlation of alarm data can represent the real relation more accurately.

Example 1: \( x \) is a process variable with standard Gaussian distribution in normal state, and in abnormal state it follows a Gaussian distribution with mean of 4 and standard deviation of 0.5. \( x \) has a probability of 0.2 to be abnormal and 0.8 to be normal. The time trend and the histogram of \( x \) are shown in

\[
y = x + 0.5v, \tag{9}
\]

where \( v \) is a random variable with standard Gaussian distribution, shown in Fig. 2(c) and (d). It is observed that \( x \) and \( y \) are strongly correlated, whose correlation coefficient is 0.96. But if the alarm limit of \( x \) and \( y \) varies from 0 to 4 individually, the corresponding correlation coefficients varies and forms a surface shown as Fig. 3. The maximum is 0.96 reached at \((2.7, 2.5)\). Compared with Fig. 2, the alarm limits corresponding to the maximum correlation between \( x \) and \( y \) are within the valleys of their respective probability density functions. Thus the practical method to choose the alarm limit can be regarded as an optimization problem based on statistical property:

\[
\text{Min}_{x_0,y_0} |\rho_{x_0,y_0} - \rho_{x,y}|. \tag{10}
\]
The above result is consistent with the basic strategy of limit choice. As shown in Fig. 1, the distribution in normal state is the left solid curve and the distribution in abnormal state is the right one. The state change is considered as a strict switching. The probability of false alarm \( p_f \) is the right side area of \( x_0 \) under the left curve, while the probability of missed alarm \( p_M \) is the left-hand side area of \( x_0 \) under the right curve. If the limit is set higher, then \( p_M \) decreases along with the increase of \( p_f \). However, if the limit is set lower, then \( p_f \) decreases along with the increase of \( p_M \). The relationship between \( p_M \) and \( p_f \) can be shown as a ROC (receiver operating characteristic) curve (shown in Fig. 4) [7] (slightly different from the normal ROC where \( 1-p_F \) is plotted versus \( p_M \)). The best limit is chosen according to an optimization:

\[
\text{Min} \left(\alpha p_M + (1-\alpha)p_F \right), \quad (11)
\]

where \( \alpha \) is a weighting factor subjectively chosen to reflect the degree of importance. The choice of the value of \( \alpha \) entails a difficult value judgment of the relative importance of false alarm and missed detection. One has to find out an acceptable trade-off between \( p_M \) and \( p_f \), which is typically the tangential point of ROC curve and one of the lines with slope \( \alpha/(1-\alpha) \). Usually, as in the previous example, the best choice of limit is within the valley of the probability density function curve.

However, this is just a conceptual explanation but not a practical method because the true normal/abnormal state is unknown thus the probabilities are unknown. Moreover, the switching between normal and abnormal state is usually fuzzy and thus hard to identify from the probability density function curve. The optimization method should then be used, as in the next example.

**Example 2:** \( x \) has a standard Gaussian distribution, and \( y \) is also determined by (9). Thus both the variables have Gaussian distribution. Based on the above method, the optimal value of \( x_0 \) and \( y_0 \) are 2.8 and 3.0 according to the correlation surface shown in Fig.5. The correlation of alarm data is close to the one of process data, 0.89.
\[ \phi_n(k) = E[x_iy_{i+k}], \quad k = -n + 1, \ldots, n - 1. \]  
(12)

Then the maximum time delayed correlation is
\[ \rho = \max \{\phi_{\max}, -\phi_{\min}\}, \]  
(13)

where
\[ \phi_{\max} = \max_k \{\phi_n(k), 0\} \geq 0, \]  
(14)
\[ \phi_{\min} = \min_k \{\phi_n(k), 0\} \leq 0. \]  
(15)

The method for limit optimization should be based on this estimated time delayed correlation.

IV. ALARM LIMIT OPTIMIZATION BY CORRELATION ANALYSIS IN AN INDUSTRIAL PROCESS

In an industrial process, large amount of process variables are connected by a causal network, in which the variables are interconnected with directions. The above method is suitable for a pair of variables to find their optimal limits. But in an industrial process, variables form a network and a variable may be connected to several variables, hence the optimal limits obtained with respect to several links are inconsistent, which often results in confusion. However, the reliability of correlation information for each link is different, some of which can be screened by specific tests. In addition, the correlation information that passes the tests can also be quantified as an index. Based on this index, a method can be developed to determine a sequence of links, then the above method can be applied one by one. For this purpose, a network should be constructed first based on process data.

According to (12) through (15), the time delayed correlation coefficients for each pair of variables with the directions (time delays) are estimated. From (14) the argument to reach the maximum is \( k_{\max} \), and from (15) the argument to reach the minimum is \( k_{\min} \). The estimated time delay is:
\[ \lambda = \begin{cases} 
\lambda_{\max}, & \phi_{\max} \geq -\phi_{\min}, \\
\lambda_{\min}, & \phi_{\max} < -\phi_{\min}.
\end{cases} \]  
(16)

Note that \( \lambda < 0 \) means the delay is from \( y \) to \( x \). These time delays between every two variables are arranged in a causality matrix \( \Lambda \). If the number of variables is large, then the order should be rearranged in a specifically ranked order, i.e. from upstream to downstream according to time delay, to make the order consistent with the propagation path, if no cycle exists in the network composed of these variables. They are arranged in a correlation matrix \( P \), the variable order (i.e. column/row order) of which is determined by a matrix transformation method [9] or a graph operation method [10].

There is likely much inaccuracy and errors in these estimates, which requires significance tests to delete the improper ones. Bauer and Thornhill [11] presented some significance testing methods for the correlation analysis, and based on them, the propagation paths can be constructed to form a causal map. We summarize the main results with slight modifications as follows. The three tests are:

1) Correlation test
   Every two variables in pair have some correlation computed by the estimate of (2), even if they are uncorrelated in nature. That is because of the measurement noise and the estimation error. If the maximum correlation \( \rho \) is smaller than a threshold, then the correlation between these two variables is ignored.

2) Directionality test
   In order to confirm the directionality, a directionality index has been introduced (Bauer and Thornhill, 2008). Here we revise it as
\[ \psi = \frac{\phi_{\max} + \phi_{\min}}{\phi_{\max} - \phi_{\min}} \in [-1,1]. \]  
(17)

If this index is too small in magnitude, then the directionality cannot be confirmed.

3) Consistency test
   The propagation of time delay should follow the superposition property, i.e. the time delay from \( x \) to \( y \) should be equal to the summation of time delays from \( x \) to \( z \) and \( z \) to \( y \). One could check this in the causality matrix. If the deviation is too large, then the check is unsatisfied. Each variable is evolved in \( p - 2 \) equations, where \( p \) is the number of variables. The numbers of checks fulfilled are recorded and arranged in a validation matrix \( \Lambda^* \).

With the aid of the above three tests, some entries in the causality matrix are replaced by empty cells. Then one can transform it into a causal map that is a digraph to represent the system topology of causality. In Bauer et al’s paper [9], one of the two typical topologies is generated according to the number of nonzero entries in the first row and above the main diagonal. Topology I is a serial arrangement of variables (likes a string of beads, as shown in Fig. 6(a)), while Topology II includes one root cause connected to all the other variables (likes an umbrella, as shown in Fig. 6(b)). Of course, a real topology is a combination of these two.

For Topology I, the link from \( x_1 \) to \( x_2 \) is considered first to optimize the limits of \( x_1 \) and \( x_2 \). Then \( x_2 \) and \( x_3 \) is considered to optimize the limit of \( x_2 \), with the limit of \( x_1 \) determined earlier.

Fig. 6. Typical types of the generated causal map topology. (a) Topology I; (b) Topology II.
For Topology II, the links are considered in the order of time delayed maximum correlation. In these two topologies, the limit of $x_1$ has been determined in the first step, thus one limit is to be determined in each following step.

V. CONCLUSION

Because of the binary format of alarm data, there is discrepancy between correlation of process data and alarm data. The alarm data is generated according to alarm limits. If the limits change, the resulting correlations also change. This property is used to optimize the alarm limits to obtain the closest correlation representation.

The alarm limits referred to in this paper are based on correlation analysis, but are not based on the safety issues. In real applications, an alarm limit obtained by this method can be used as one of the limits because it reflects the state or condition transition including normal/abnormal transition. In most cases, the former is stricter than the latter, so it can be regarded as a reference. However the alarm limits based on safety issues should generally remain, or be changed cautiously after careful analysis.

The correlation analysis is between each pair of variables, thus the correlation color map is used for an industrial process. However, the correlation considering the time delay is an estimate whose property is not accurate. This is the reason why various consistency checks are made for the estimated correlations and time delays. The entries with less reliability are ignored. These methods are not perfect and special manual checks are often necessary.

In this research, a case study of a consolidated tailings process at Suncor Energy Inc. is made based on industrial data. Causalities between three variables have been validated by process data and the optimized limits have been validated by process knowledge and experience. As in Fig. 7(a), the limit of $x_2$ can differentiate two system modes clearly. In Fig 7(b), the distributions of the two system modes are both approximately Gaussian and they have apparent overlap. Although the limit of $x_4$ cannot totally differentiate them, it can somehow compromise the missed alarms and false alarms because it is in the valley between the two peaks. The case study has proved that the proposed method is reasonable and practical to some extent although rigorous theoretical derivation is difficult to obtain. It is to be noted that causality is unequal to correlation, thus the causal map obtained from process data cannot be validated without process knowledge. Connectivity and topology in large-scale systems should be used as a priori process knowledge. In this aspect, a signed digraph may be a useful tool [10].

Another issue to be noted is that the alarms in this paper refer to deviation alarms that are associated directly with the variable values. Other types of alarms such as bad measurement alarms, systems diagnostic alarms and instrument diagnostic alarms [2] have other properties and their correlations may not be computed by using (1) and (2).

ACKNOWLEDGMENT

Data and support from Suncor Energy Inc. is gratefully acknowledged.

REFERENCES