Journal of Loss Prevention in the Process Industries 45 (2017) 43–55

Contents lists available at ScienceDirect

Journal of Loss Prevention in the Process Industries

journal homepage: www.elsevier.com/locate/jlp

Abnormal data detection for multivariate alarm systems based on correlation directions☆

Yan Yu [a], Di Zhu [a, b], Jiandong Wang [c, *], Yan Zhao [d]

[a] College of Engineering, Peking University, Beijing, China
[b] Dept. of Aeronautical and Mechanical Engr., PLA Aviation Institute, Beijing, China
[c] Shandong University of Science and Technology, Qingdao, Shandong Province, China
[d] Shandong Electric Power Research Institute, Jinan, Shandong Province, China

A R T I C L E I N F O

Article history:
Received 3 June 2016
Received in revised form
25 November 2016
Accepted 25 November 2016
Available online 29 November 2016

Keywords:
Industrial alarm systems
Abnormal data segments
Correlation direction
Piecewise linear representation

A B S T R A C T

This paper proposes a method to detect abnormal data segments from historical multivariate time series, which are common prerequisites for rationalization of industrial alarm systems. Correlation directions among process variables are taken as the features to detect abnormal conditions. To find time instants of changing correlation directions, key turning points (KTPs) are determined by a piecewise linear representation of multivariate time series. Correlation directions in each data segment between adjacent KTPs are calculated from Spearman's rank correlation coefficients and associated hypothesis tests. Data segments are classified into normal or abnormal ones by comparing the calculated correlation directions with their counterparts in normal conditions obtained from process knowledge. Numerical and industrial examples are provided to illustrate the proposed method.

1. Introduction

Industrial alarm systems play an important role as safeguards to prevent abnormal conditions having negative effects on the safe and efficient operations of modern industrial plants (Kim, 1994; Bransby and Jenkinson, 1998; Stauffer and Clarke, 2016). Industrial alarm systems monitor hundreds and thousands of process variables in a real-time manner. They are the indispensable computerized tools for industrial plant operators to promptly detect the occurrences of abnormal conditions, and to properly take corrective actions. Retrospective investigation on a large number of accidents reveals that almost every accident is accompanied with one or several abnormal conditions occurred in advanced as precursors (Macdonald, 2004; Pariyani et al., 2010; Beebe et al., 2013); in other words, these accidents could be avoided if the accompanied abnormal conditions were detected promptly and handled properly. However, many existing industrial alarm systems suffer from a problem of alarm overloading, namely, there are too many alarms to be promptly handled by industrial plant operators. One main cause of alarm overloading is that the design of industrial alarm systems is isolated from related process variables (Wang et al., 2016).

In order to design more effective alarm systems based on multiple process variables, a common prerequisite is the availability of training data sets with labelled normal and abnormal data samples, from which some features are extracted in order to detect abnormal conditions in an online manner. Brooks et al. (2004) developed dynamic alarm thresholds for multiple process variables from their best operating zones visualized in parallel coordinates via a geometric process control technique. Gupta et al. (2013) rationalized alarm thresholds for multiple process variables to alleviate noise effects and detect faults earlier by an integration of techniques such as wavelet analysis and principal component analysis. Alrowaie et al. (2014) designed alarm thresholds based on particle filters for multivariate nonlinear stochastic systems. Zhu et al. (2014) introduced a method based on Bayesian regression for designing dynamic alarm thresholds for different operating stages. Zang and Li (2014) and Han et al. (2016) optimized alarm thresholds for multiple process variables by minimizing false and missed alarm probabilities based on joint probability densities of process variables in the normal and abnormal conditions. However, the above studies assumed that normal and abnormal data samples or their statistics such as joint probability densities were available at hands, without discussing how to obtain them from historical data samples. On one

☆ This research was supported by the National Natural Science Foundation of China under grant No. 61433001, and the Research Fund for the Taishan Scholar Project of Shandong Province of China.

* Corresponding author.
E-mail addresses: yuyan_coe@pku.edu.cn (Y. Yu), zhudi@pku.edu.cn (D. Zhu), jiandong@sdust.edu.cn (J. Wang), zhaoyan@yjy.sd.sgcc.com.cn (Y. Zhao).

http://dx.doi.org/10.1016/j.jlp.2016.11.011
0950-4230/

hand, manually distinguishing normal and abnormal data segments is possible only for small-sized data sets; on the other hand, it is generally true that more training data samples are beneficial to improve the performance of designed multivariate alarm systems. Therefore, it is very necessary to have an automatic method to detect abnormal data segments from historical data samples.

Our principle to detect abnormal data segments is to look at the correlation directions among related process variables. Let us take a feedwater pump in a large-scale thermal power plant (to be further studied later in Section 5) as an example. Based on some preliminary physical laws of pumps, it is ready to know a fact that the outlet flow of the feedwater pump is always increasing or decreasing in a synchronized manner with the pump speed in the normal conditions, and such a relation no longer holds when abnormal conditions occur. According to this fact, analyzing the correlation direction of related process variables is capable of detecting abnormal data segments. This is related to the research topics of qualitative trend analysis (QTA) and piecewise linear representation (PLR) in the field of time series data mining.

The main task of QTA is to extract trends by segmenting signals into non-overlapping episodes or representing the process trends using the predefined "basic shapes", e.g., triangular episode (Cheung and Stephanopoulos, 1990) or increasing, decreasing and steady primitives (Janusz and Venkatasubramanian, 1991; Charbonnier et al., 2005); see Zhou and Ye (2016) for a comprehensive study of the polynomial fit-based QTA algorithms.

The main goal of the PLR is to segment and approximate original time series in the form of straight lines. Thus, the PLR is a useful tool for data preprocessing before the step of QTA. Duncan and Bryant (1996) suggested a dynamic programming approach to achieve the global optimal solution for PLR, which however is hard to be applied in practice due to excessive quadratic running time. Keogh et al. (2004) categorized the segmentation algorithms based on piecewise linear approximation into three groups, namely, top-down, bottom-up and sliding-windows approaches. Within the three groups of approaches, the choices of PLR techniques to be exploited in this paper are manifold. One interesting technique aims at identifying some prominent points to be consistent with human intuition and episodic memory (Perng et al., 2000). Major works on this technique are as follows: Pratt and Fink (2002) proposed a PLR algorithm based on discarding minor fluctuations and keeping major minima and maxima for fast compression of time series. Chung et al. (2004) and Fu et al. (2007) presented a PLR method based on searching for perceptually important points (PIPs), and showed the effectiveness of this method via real-time applications.

This paper proposes a method to detect abnormal data segments from historical multivariate time series. Correlation directions among process variables are taken as the features to detect abnormal conditions. So-called key turning points (KTPs) are determined by finding the PLR of multivariate time series, to be associated with time instants of changing correlation directions. Correlation directions for each data segment between adjacent KTPs are calculated from Spearman's rank correlation coefficients and associated hypothesis tests. Data segments are classified into normal or abnormal ones by comparing the calculated correlation directions with their counterparts in normal conditions obtained from process knowledge. The novelties of the proposed method mainly lie in choosing correlation directions between process variables as the features for abnormal data detection, and extending the PLR method based on PIPs proposed by Chung et al. (2004) and Fu et al. (2007) from univariate time series to multivariate ones in order to determine the KTPs.

The rest of this paper are organized as follows. The problem to be solved is described in Section 2. Section 3 introduces the basic idea of the proposed method via an example. The details of the proposed method is presented in Section 4. Section 5 provides numerical and industrial examples to illustrate the proposed method. Section 6 concludes the paper.

2. Problem description

Consider a multivariate system involving m process variables $X := [X_1, X_2, \ldots, X_m]$ and denote the corresponding multivariate time series T composed by X as

$$T := \{X_i(n)\}, \ i = 1, \ldots, m, \ n = 1, \ldots, N,$$

where n and N are the sampling time index and the number of data samples, respectively. Only related variables are chosen based on some preliminary process knowledge to be present in X. A correlation direction between $X_i \in X$ and $X_j \in X$ in a segment of T, denoted by the symbol 's', is defined as

$$\text{sign}_s(X_i, X_j) = \begin{cases} 1, & \rho_s(X_i, X_j) > 0, \\ 0, & \rho_s(X_i, X_j) = 0, \\ -1, & \rho_s(X_i, X_j) < 0. \end{cases} \quad (1)$$

Here $\rho_s(X_i, X_j)$ represents a certain type of correlation coefficient between X_i and X_j in the segment s. The value 1, −1, or 0 of $\text{sign}_s(X_i, X_j)$ indicates a positive, negative or insignificant correlated relation between X_i and X_j.

Assume that the values of $\text{sign}_s(X_i, X_j)$ for all pairs of X_i and X_j are constant in normal conditions, denoted as $\text{sign}(X_i, X_j)$. From some process knowledge of X, the values of $\text{sign}(X_i, X_j)$ in normal conditions are known a prior.

A classification rule for abnormal data segments is to classify the data samples in a segment s as being abnormal, if there exists a pair of variables X_i and X_j such that $\text{sign}_s(X_i, X_j)$ is inconsistent with the corresponding values of $\text{sign}(X_i, X_j)$ in normal conditions.

In this paper, our objective is to divide a given historical data sequence T of X into multiple segments in a proper manner, calculate correlation directions among all process variables in the segments, and identify abnormal data segments according to the above classification rule.

3. Basic idea

This section introduces the basic idea of the proposed method to solve the problem described in Section 2 via an example.

The most critical step of the proposed method is to divide a given historical data sequence into segments according to the change of correlation directions among process variables. It is ready to see that a point in a data sequence with the maximum distance to a line connecting the starting and ending points of the sequence has the greatest impact on the correlation direction. Thus, we need to detect some special points, referred to as key turning points (KTPs), representing the changing points of correlation directions among process variables. Fig. 1 shows a bivariate time series as an example to explain the basic steps of the proposed method to determine KTPs.

First, as shown in Fig. 1-(a), the time sequence with the starting point A at $n = 1$ and ending point B at $n = 40$ is considered as one segment, in which A and B are KTPs by default. Second, the orthogonal distance from each point inside the current segment AB to the corresponding line connecting two adjacent KTPs A and B is calculated to find the point with the maximum distance as a KTP. Here the point C at $n = 20$ with the maximum orthogonal distance is detected as the third KTP so that there are two segments AC and CB. Next, by recalculating the orthogonal distances of all points in

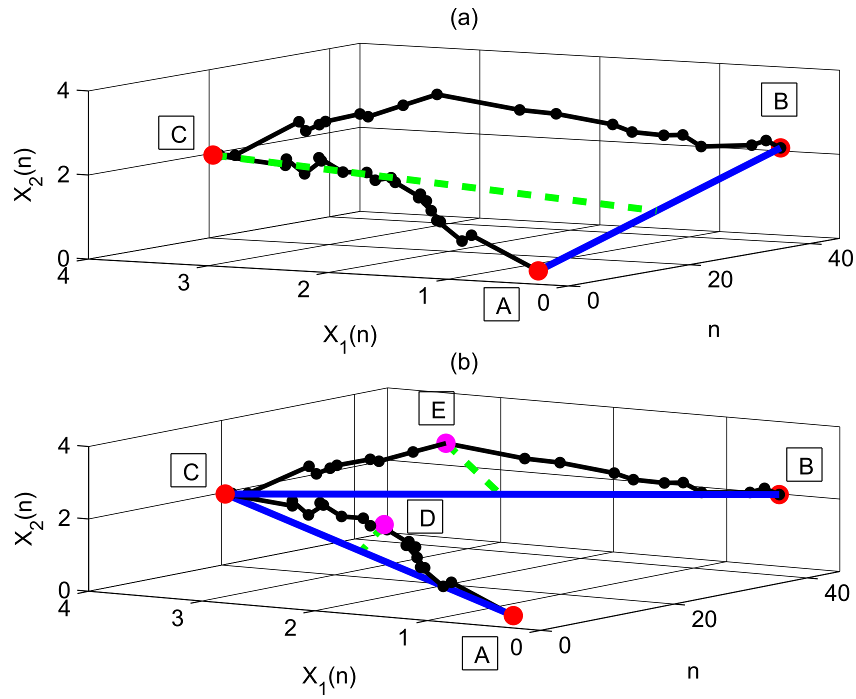

Fig. 1. The schematic diagram of determination of KTPs.

each segment and choosing the point with the maximum orthogonal distance of all segments, new KTPs can be obtained successively. The procedures of determining the fourth KTP E at $n = 30$ are shown in Fig. 1-(b), where the maximum orthogonal distance from point E in the segment CB is larger than that from point D in the segment AC.

There is an important issue to be addressed in the determination of KTPs. Owing to the presence of noises, the time interval between two adjacent KTPs could be very short, namely, there are only few points in a segment. In general, the relations among process variables within a short period of time are regarded to be fixed. Thus, a parameter δ is introduced to as the minimum time interval between adjacent KTPs. The rationale of δ is the same as alarm delay timers, which raise an alarm only when several consecutive samples overpass an alarm threshold. Thus, δ is taken as the stop condition for the iterations in determining KTPs, i.e., the iterations stop if choosing any new point as a KTP cannot meet the requirement on δ.

The basic idea of the proposed method is summarized as follows. First, the multivariate time series is divided into segments by the above-mentioned determination of KTPs. Second, correlation directions among process variables in data segments are calculated by performing a certain type correlation analysis on each pair of process variables and its associated hypothesis tests. Finally, by comparing the calculated correlation directions with their counterparts in normal conditions, the data segments are classified into normal and abnormal ones.

4. The proposed method

This section presents the detailed steps of the proposed method.

4.1. Determination of KTPs

First, in order to prevent that some process variables with large amplitudes erroneously play dominate roles, the proposed method preprocesses the raw data by normalizing the observations of each process variable in order to have zero mean and unit variance. Given the raw data of a multivariate time series $\tilde{T} := \{\widetilde{X}_i(n)\}$ for $i = 1, \ldots, m$ and $n = 1, \ldots, N$, the normalized time series T is obtained as

$$T := \{X_i(n)\} := \left\{ \frac{\widetilde{X}_i(n) - \overline{\widetilde{X}}_i}{\sigma_{\widetilde{X}_i}} \right\},$$

where $\overline{\widetilde{X}}_i$ and $\sigma_{\widetilde{X}_i}$ denote the sample mean and standard deviation of $\widetilde{X}_i$, respectively.

Second, the segmentation based on the determination of KTPs is mathematically defined. The sequence T can be divided into K non-overlapping and contiguous segments by using $(K+1)$ KTPs. The sampling time indices of the KTPs are denoted as $S := \{p_1, \ldots, p_{K+1}\}$. Two special indices are $p_1 = 1$ and $p_{K+1} = N$. The j-th segment s_j for $j = 1, \ldots, K$ is defined as the subsequence $s_j := \{X_i(n)\}$ for $p_j \leq n \leq p_{j+1}$.

Next, for the $m+1$ dimensional linear space composed of m process variables $X_1, X_2, \ldots, X_m$ and the time index n, the calculation of the orthogonal distance from a point to a line is calculated as follows, according to the vector geometry theory (Schneider and Eberly, 2002). The orthogonal distance from a point $P := [X_{1P}, X_{2P}, \ldots, X_{iP}, n_P]$ to the line AB between the point $A := [X_{1A}, X_{2A}, \ldots, X_{iA}, n_A]$ and $B = [X_{1B}, X_{2B}, \ldots, X_{iB}, n_B]$ is defined as the minimum Euclidean distance from P to any point P_0 located on AB, namely, the distance between P and its projection on AB. The line AB is mathematically described as

$$\frac{X_1 - X_{1A}}{X_{1B} - X_{1A}} = \cdots = \frac{X_i - X_{iA}}{X_{iB} - X_{iA}} = \frac{n - n_A}{n_B - n_A} = \beta, \;\; i = 1, \ldots, m, \qquad (2)$$

where β is an unknown parameter to be determined. Eq. (2) says that the coordinate of P_0 can be represented as

$[(X_{iB}-X_{iA})\beta+X_{iA},(n_B-n_A)\beta+n_A]$ for $i=1,\ \ldots,\ m$. Let $\left\|\overrightarrow{P_0P}\right\|_2$ denote the distance between P and P_0 (or equivalently the 2-norm of the vector $\overrightarrow{P_0P}$) as

$$\begin{aligned}\left\|\overrightarrow{P_0P}\right\|_2 &= \sqrt{[(X_{iB}-X_{iA})\beta+X_{iA}-X_{iP}]^2+[(n_B-n_A)\beta+n_A-n_P]^2}\\ &= \sqrt{\left(\left\|\overrightarrow{AB}\right\|_2\right)^2\beta^2+2\left(\overrightarrow{AB}\cdot\overrightarrow{AP}\right)\beta+\left(\left\|\overrightarrow{AP}\right\|_2\right)^2},\end{aligned} \tag{3}$$

where

$$\begin{aligned}\overrightarrow{AB} &= [(X_{iB}-X_{iA}),(n_B-n_A)],\ \ i=1,\ \ldots,\ m,\\ \overrightarrow{AP} &= [(X_{iP}-X_{iA}),(n_P-n_A)],\ \ i=1,\ \ldots,\ m.\end{aligned}$$

The orthogonal distance is the one to minimize $\left\|\overrightarrow{P_0P}\right\|_2$. Set the first derivative of $\left\|\overrightarrow{P_0P}\right\|_2^2$ with respect to β to zero and solve the corresponding equation to reach the optimal solution $\beta^*=-(\overrightarrow{AB}\cdot\overrightarrow{AP})$. By inserting β^* into (3), the orthogonal distance D is obtained,

$$D=\frac{\sqrt{\left(\left\|\overrightarrow{AP}\right\|_2\left\|\overrightarrow{AB}\right\|_2\right)^2-\left(\overrightarrow{AB}\cdot\overrightarrow{AP}\right)^2}}{\left\|\overrightarrow{AB}\right\|_2}. \tag{4}$$

Finally, based on the above steps, the KTPs can be picked up one by one in a recursive manner, as given by the pseudo codes in Algorithm 1. The determination of KTPs essentially stops in the case that adding one more KTP will break the inequality $\min(z_j)\geq\delta$, where $z_j=p_{j+1}-p_j$ denotes the number of data points in the j-th segment, and δ is the minimum time interval to avoid obtaining too short segments caused by noises. Algorithm 1 extends the PLR method based on PIPs proposed by Chung et al. (2004) and Fu et al. (2007) from univariate time series to multivariate ones. Since the PLR method is a top-down segmentation approach, its computational cost is low as analyzed in Chung et al. (2004) and Fu et al. (2007). Thus, the computation cost of Algorithm 1 is minor, too. In particular, the time series T is divided into K non-overlapping and contiguous segments, and in each iteration to determine a new segment, the orthogonal distances from N points to a line between associated two KTPs are calculated; thus, the computation complexity is $O(KmN)$. In the worst situation, the segment number is $K=N/\delta$ and the computational complexity of Algorithm 1 is $O(\mathrm{mN}^2/\delta)$.

Algorithm 1.

$Input$: normalized multivariate time series T, minimum interval δ
$Output$: KTPs in the detected order Q
$Q=Q_r=[1,N]$ ▷ The starting and ending points are both KTPs
for $k=3:N$ **do**
  $D=zeros(1,N)$
  for $i=2:length(Q_r)$ **do**
    $p_s=Q_r(i-1)+\delta$
    $p_e=Q_r(i)-\delta$ ▷ The range for searching is limited by δ
    for $j=p_s:p_e$ **do**
      Calculate $D(j)$ for the j-th point by Eq. (4)
    end for
  end for
  if $\max(D)>0$ **then**
    $Q(k)=$ time index of $\max(D)$
    $Q_r=Q$ in the ascending order
  else
    Stop the algorithm and obtain the output Q
  end if
end for

4.2. Choosing the number of segments

The piecewise linear representation (PLR) of the normalized multivariate series using the obtained KTPs limited by δ could be over-fitted due to the noise effect. Thus, in order to reduce the over-fitting errors, it is necessary to select an appropriate number K of segments.

Assume that the multivariate time series T is divided by $K+1$ KTPs, $p_1, \cdots, p_{K+1}$, where $p_1 = 1$ and $p_{K+1} = N$. The PLR of T is denoted as:

$$T_{PLR} = \langle f_1[(X_i(p_1), p_1), (X_i(p_2), p_2)], \ \ldots, \ f_K[\times (X_i(p_K), p_K), (X_i(p_{K+1}), p_{K+1})] \rangle$$

where $f_j[(X_i(p_j), p_j), (X_i(p_{j+1}), p_{j+1})]$ represents the linear fitting function in the segment $[p_j, p_{j+1}]$. The fitting function can be obtained by a linear regression. Mathematically, the estimates $(\widehat{X}_i(n), n)$ for $i = 1, \ \ldots, \ m$ and $n = 1, \ \ldots, \ N$ can be understood as the projection of the data point $(X_i(n), n)$ on the line connecting the starting and ending points of the corresponding segment. The fitting error is defined as

$$E = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\left\| (X_i(n), n) - \left(\widehat{X}_i(n), n \right) \right\|_2 \right)^2}, \quad i = 1, \ \ldots, \ m.$$

The fitting error can be equivalently calculated as

$$E = \sqrt{\frac{1}{N} \sum_{n=1}^{N} D(n)^2} \tag{5}$$

where $D(n)$ denotes the orthogonal distance from the point $(X_i(n), n), i = 1, \ \ldots, \ m$ and $n = 1, \ \ldots, \ N$ to the corresponding segment. Since $X_i(n)$'s for $i = 1, \ \ldots, \ m$ are normalized to unit variance, a constant $E_{\max}$ is introduced as the upper bound of E: if E is larger than $E_{\max}$, then the fitting error is unacceptable and the results from the proposed method are not reliable. A default value of $E_{\max}$ is taken as $E_{\max} = 1$.

The fitting error E is a loss function varying with the parameter K (the number of segments). It seems infeasible to choose in advance a fixed appropriate value of K for different noise levels. Here we recommend a practical way to select K by following the parsimony principle and taking the same hypothesis test for model order determination in the field of system identification. For two segment numbers K_1 and K_2 with $K_1 < K_2$ and their associated fitting errors $E(K_1)$ and $E(K_2)$, the quantity

$$f = \frac{E(K_1) - E(K_2)}{E(K_2)} \frac{n - K_2}{K_2 - K_1} \tag{6}$$

is with F-distribution $F(K_2 - K_1, n - K_2)$ (Soderstrom and Stoica, 1989). The selection of K is implemented in the Algorithm 2.

Algorithm 2.

i. Denote the obtained KTPs by Algorithm 1 as $Q = [q_1, q_2, \ \ldots, \ q_{K_{\max}+1}]$. For $2 \le K \le K_{\max}$, calculate the loss function E in (5) as $E(K)$.
ii. Draw the scatter plot of $(K, E(K))$ for $2 \le K \le K_{\max}$. Set an appropriate value K_{opt} at a place that the decrement of $E(K)$ with the increment of K becomes small, which is statistically judged by a hypothesis test based on F-distributed statistic f in (6).
iii. Obtain the KTPs $Q_{K_{opt}} = [q_1, q_2, \ \ldots, \ q_{K_{opt+1}}]$ as the final KTPs for the subsequent calculation of correlation directions.

The segmentation by Algorithm 1 has a nice property, namely, adding a new KTP has a local effect in terms that only one of the obtained segments is partitioned without affecting other segments. Therefore, the proposed method is not sensitive to the value of K_{opt}, as long as the variations of $E(K)$ around K_{opt} are insignificant.

Time delays are common for process variables in practice. If there is a time delay between two process variables, one variable may change its direction while the other still follows the previous direction. Thus, the presence of time delays may result in a false detection of abnormal segments. Hence, the time delay between X_i and X_j needs to be estimated, and either X_i or X_j is shifted

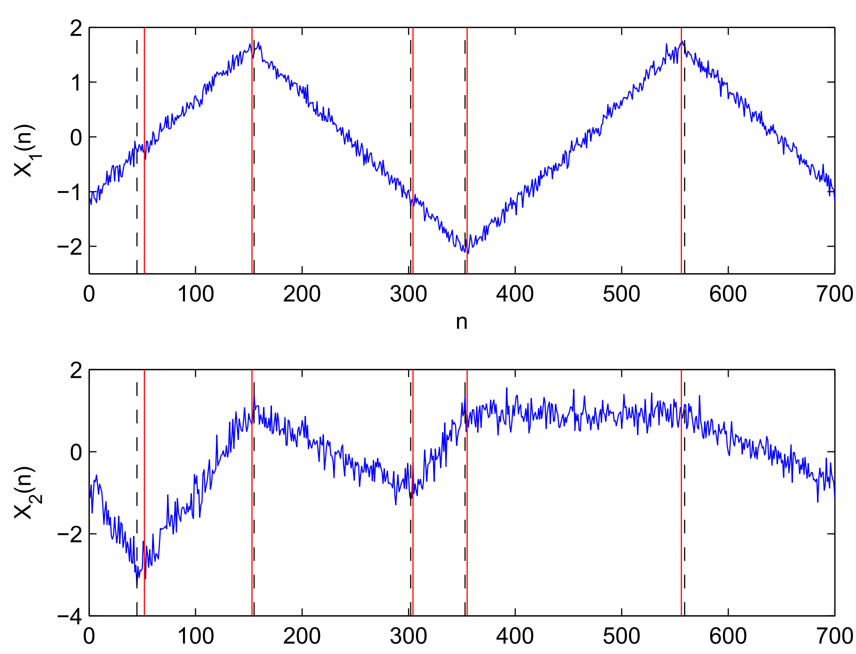

Fig. 2. The normalized time series $X_1(n)$ and $X_2(n)$ with the actual (solid vertical line) and detected (dash vertical line) changing time instants of correlation directions.

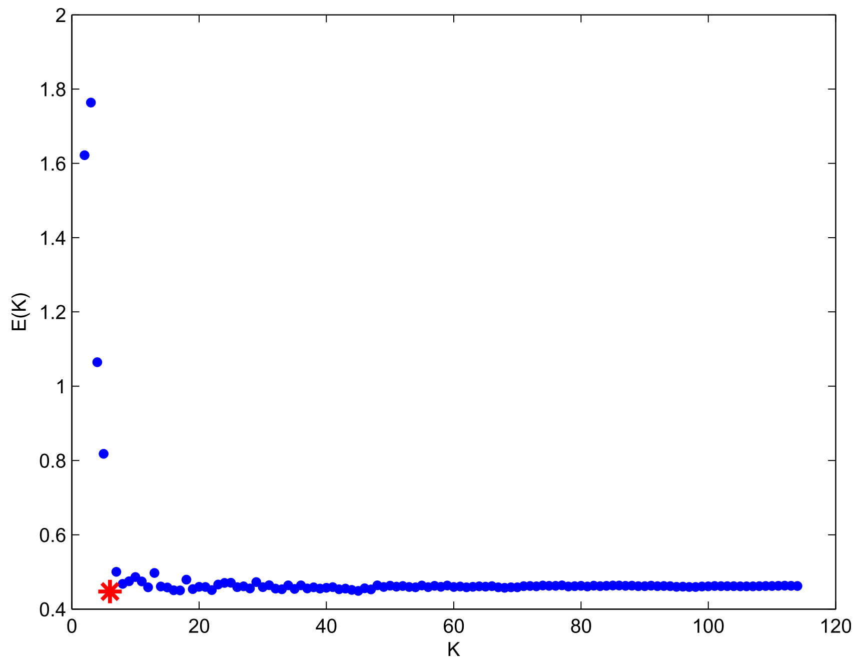

Fig. 3. The scatter plot of $E(K)$ and K with the selected point K_{opt} (star).

accordingly to remove the negative effect of time delays. As studied in a survey (Björklund, 2003), the time delay estimation has been an interesting topic in many fields, but most of the existing methods have their own limitations; in particular, experiments are usually required to introduce special signals to achieve accurate estimates of time delays. However, in this context, the historical data samples are given without special experiments so that many existing methods can not be applied. Here we propose an algorithm to estimate time delays, by taking the same principle as a common time-delay estimation method to find the maximum cross correlation coefficient between two signals. The time delay estimation is implemented in the Algorithm 3.

Algorithm 3.

i. Shift X_i by a specific time delay d_i as $X_{i,s}(n) := X_i(n+d_i)$ for $i = 1, \ldots, m$, and find the optimal number of segments from Algorithm 2 as $K_{opt}(d_1, \ldots, d_m)$ and the fitting error E in (5) as $E(K_{opt}(d_1, \ldots, d_m))$;
ii. Repeat the first step for another group of time delays, and choose the time delays as the ones to achieve the minimum fitting error, i.e.,

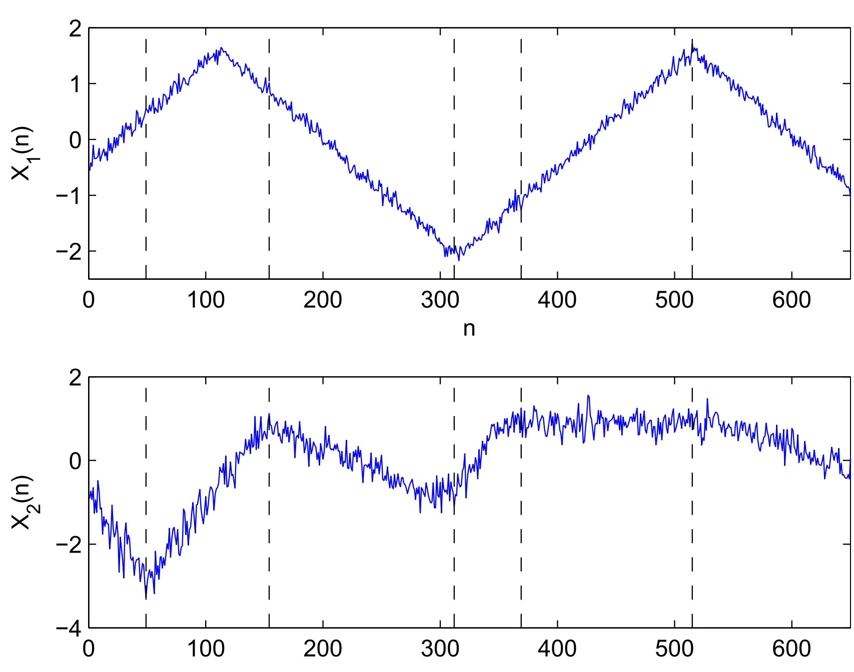

Fig. 4. The normalized time series $X_{1,\tau}(n)$ and $X_2(n)$ with the detected (dash vertical line) changing time instants of correlation directions.

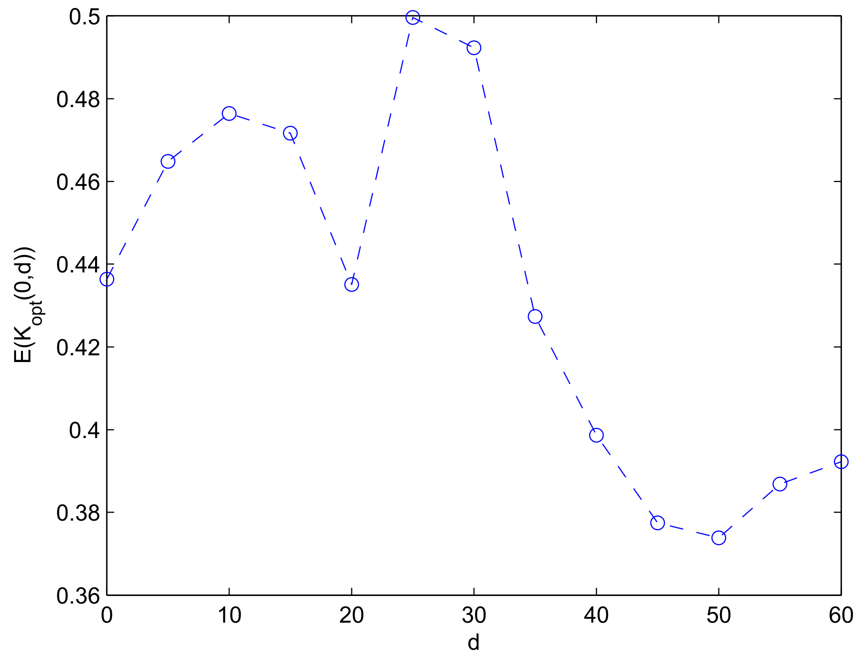

Fig. 5. The results to estimate the time delay from Algorithm 3.

$$\{d_{1,opt}, \ldots, d_{m,opt}\} = \min_{d_1 \ldots, d_m} (E(K_{opt}(d_1, \ldots, d_m)))$$

iii. Implement Algorithm 2 to $\{X_{i,s}(n) := X_i(n + d_{i,opt})\}$ for $i = 1, \ldots, m$, to obtain the KTPs for the subsequent calculation of correlation directions.

The time delay d_i can take either positive or negative values; however, some basic process knowledge is helpful to significantly reduce a search space. For instance, in Example 3 appeared later in Section 5, the generating power $X_1(n)$ in a thermal power plant follows a variation of feed-coal flow $X_2(n)$ in a short time period; thus, only positive time delays are considered for d_1.

4.3. Abnormal data classification

Given the segmentation result $Q_{K_{opt}} = [q_1, q_2, \ldots, q_{K_{opt+1}}]$, the correlation directions of the multivariate time series T can be extracted by performing correlation analysis for each data segment. Spearman's rank correlation coefficient is applied here, owing to its wide range applicability for different statistical distributions of

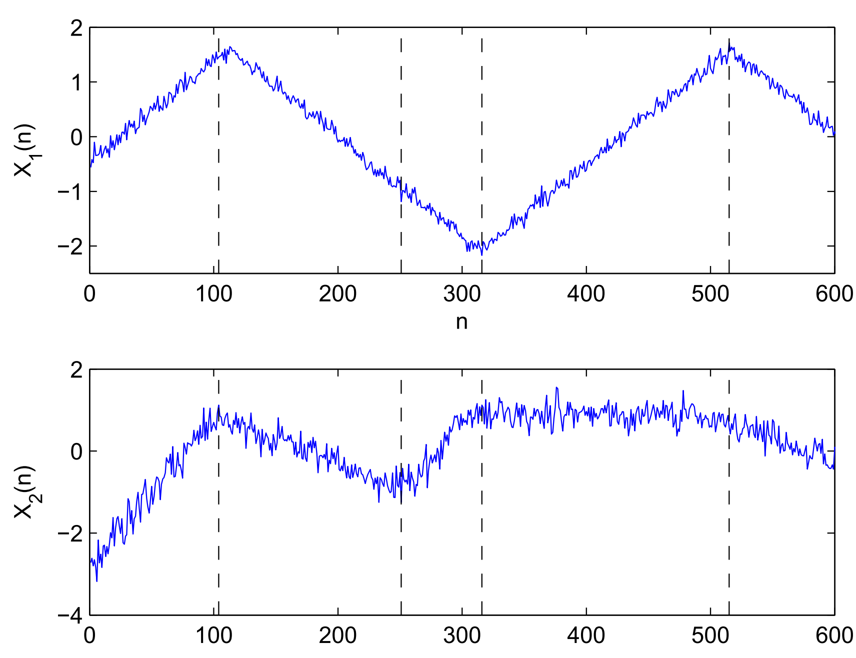

Fig. 6. The normalized time series $X_{1,\tau}(n)$ and $X_{2,d_{opt}}(n)$ with the detected (dash vertical line) changing time instants of correlation directions.

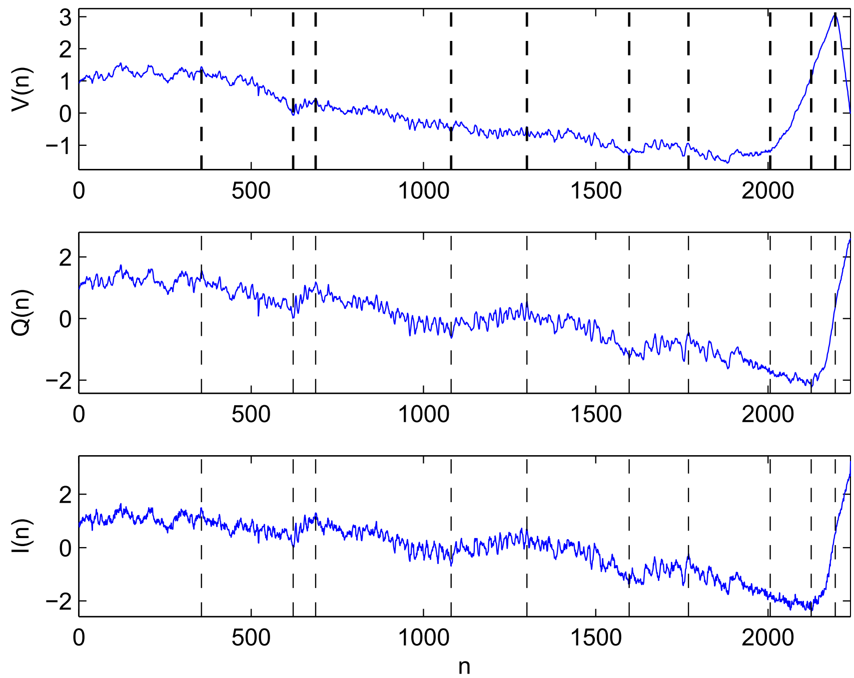

Fig. 7. The time sequence plots of $Q(n), V(n)$ and $I(n)$ with the dash vertical lines to separate segments for Example 2.

process variables. The observations of X_i and X_j in the s-th segment are ranked in an ascending order using the integers $k = 1, \ldots, z_s$ as the sample indices, where z_s represents the sample size of the s-th segment. Denote the respective ranks of X_i and X_j by $R_{k,i}$ and $R_{k,j}$ respectively. Spearman's rank correlation coefficient for X_i and X_j in the s-th segment of T is calculated by

$$\rho_s(X_i, X_j) = \frac{\sum_{k=1}^{z_s}\left(R_{k,i} - \overline{R_i}\right)\left(R_{k,j} - \overline{R_j}\right)}{\sqrt{\sum_{k=1}^{z_s}\left(R_{k,i} - \overline{R_i}\right)^2 \sum_{k=1}^{z_s}\left(R_{k,i} - \overline{R_i}\right)^2}},$$

where $\overline{R_i} := \frac{1}{z_s}\sum_{k=1}^{z_s} R_{k,i}$ and $\overline{R_j} := \frac{1}{z_s}\sum_{k=1}^{z_s} R_{k,j}$. Then it is necessary to perform one-tailed hypothesis tests for positive and negative correlation directions as

$$H_0 : \rho_s(X_i, X_j) = 0 \quad vs \quad H_1 : \rho_s(X_i, X_j) > 0,$$
$$H_0 : \rho_s(X_i, X_j) = 0 \quad vs \quad H_2 : \rho_s(X_i, X_j) < 0.$$

The hypothesis tests are based on the statistic

$$U = \rho_s(X_i, X_j)\sqrt{\frac{z_s - 2}{1 - (\rho_s(X_i, X_j))^2}},$$

which approximately follows the t distribution with $z_s - 2$ degrees of freedom (Press et al., 2007). Given a significance level α, e.g., $\alpha = 0.05$, the null hypothesis H_0 is rejected and H_1 is accepted if $U > t_\alpha(z_s - 2)$, or H_2 is accepted if $U < -t_\alpha(z_s - 2)$. Here $t_\alpha(z_s - 2)$ denotes the critical value of U. Accordingly, the correlation between X_i and X_j in the s-th segment is considered to be significant and the correlation direction $\text{sign}_s(X_i, X_j)$ in (1) takes the value of 1 or −1.

According to the classification rule in Section 2, the s-th segment is classified as being abnormal if the correlation direction $\text{sign}_s(X_i, X_j)$ between X_i and X_j in the s-th segment is inconsistent with the counterpart $\text{sign}(X_i, X_j)$ in normal conditions; otherwise, the s-th segment is classified as being normal in the case that $\text{sign}_s(X_i, X_j)$ is consistent with $\text{sign}(X_i, X_j)$ for all values of i and j.

5. Examples

This section provides numerical and industrial examples to illustrate the effectiveness of the proposed method.

Example 1. A bivariate time series $\tilde{T} = \{\widetilde{X_1}(n), \widetilde{X_2}(n)\}$ for $n = 1, \ldots, N$ is generated by concatenating six piecewise linear functions $\tilde{X}_{2i} = f_i(\tilde{X}_{1i}) + \varepsilon_i$ for $i = 1, \ldots, 6$, namely,

$$\begin{aligned}
&\tilde{X}_{11} = 0 : 0.01 : 0.5 + \varepsilon(0, 0.05), \tilde{X}_{21} = -\sqrt{3}\tilde{X}_{11} + \varepsilon(0, 0.1),\\
&\tilde{X}_{12} = 0.5 : 0.01 : 1.5 + \varepsilon(0, 0.05), \tilde{X}_{22} = \sqrt{3}\left(\tilde{X}_{12} - 1\right) + \varepsilon(0, 0.1),\\
&\tilde{X}_{13} = 1.5 - (0 : 0.01 : 1.5) + \varepsilon(0, 0.05), \tilde{X}_{23} = \tilde{X}_3 \Big/ \sqrt{13} + \varepsilon(0, 0.1),\\
&\tilde{X}_{14} = -(-0.5 : 0.01 : 0) - 0.5 + \varepsilon(0, 0.05), \tilde{X}_{24} = -\sqrt{3}\tilde{X}_{14} + \varepsilon(0, 0.1),\\
&\tilde{X}_{15} = -0.5 : 0.01 : 1.5 + \varepsilon(0, 0.05), \tilde{X}_{25} = \sqrt{3}\Big/2 + \varepsilon(0, 0.1),\\
&\tilde{X}_{16} = 1.5 - (0 : 0.01 : 1.5) + \varepsilon(0, 0.05), \tilde{X}_{26} = \tilde{X}_{16}\Big/\sqrt{3} + \varepsilon(0, 0.1).
\end{aligned} \tag{7}$$

Here $\varepsilon(0, \sigma)$ denotes Gaussian white noise with mean 0 and standard deviation σ. The time sequence plots of the normalized time series $\{X_1(n), X_2(n)\}$ for $n = 1, \ldots, 700$ are shown in Fig. 2. The correlation direction in normal conditions is pre-defined as

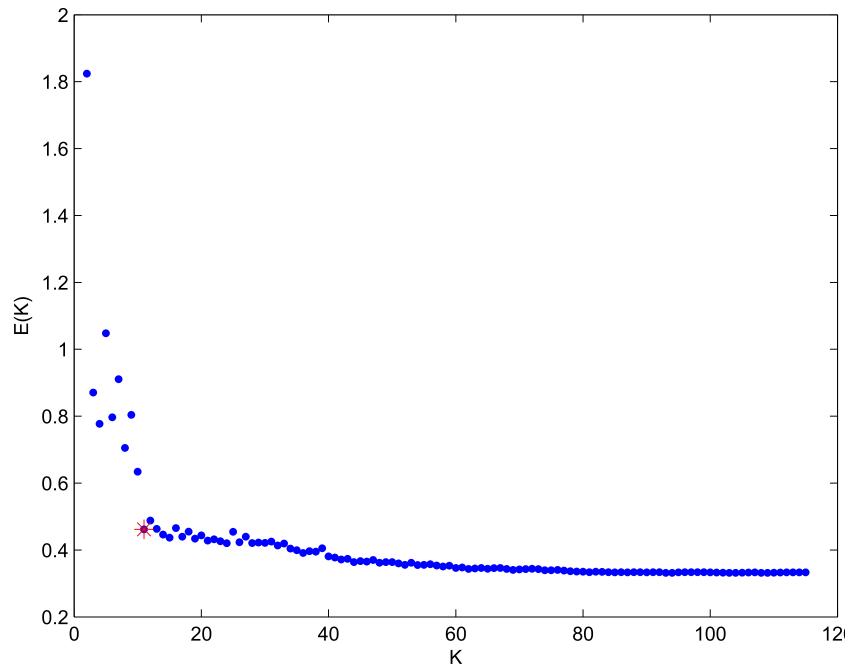

Fig. 8. The scatter plot of $E(K)$ and K with the selected point K_{opt} (star) for Example 2.

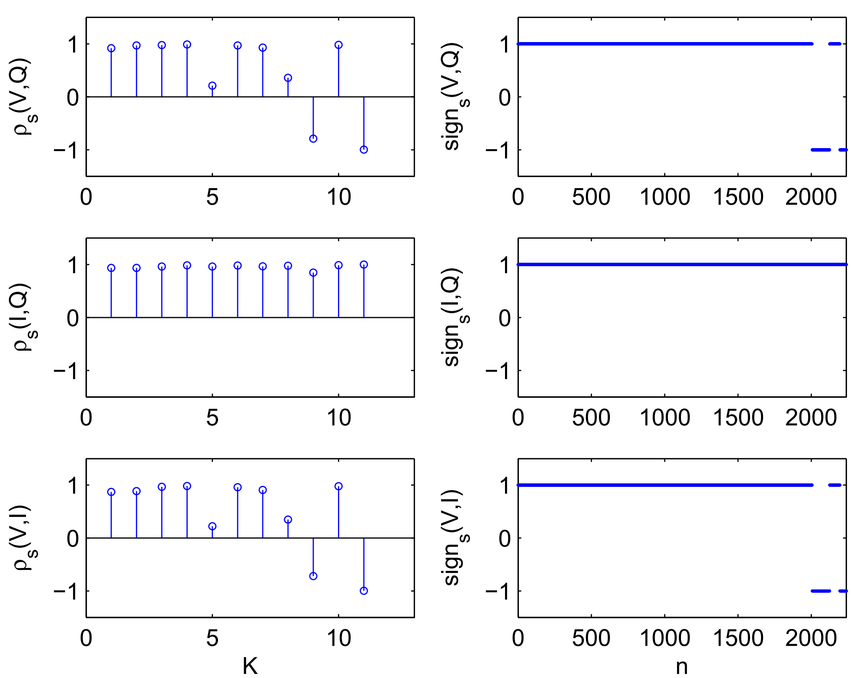

Fig. 9. The calculated correlation coefficients (left) and correlation directions (right) for Example 2.

$\text{sign}(X_1, X_2) = 1$, saying that X_1 and X_2 are in the normal condition if they increase or decrease in the same direction. It is obtained from (7) that the actual changing time instants of correlation directions between X_1 and X_2 are $n = [52, 153, 304, 355, 556]$, represented by solid (red) vertical lines in Fig. 2, and the actual correlation directions in these segment are $-1, 1, 1, -1, 0, 1$. As a result, the data samples in the 1-st, 4-th, 5-th segments should be classified as abnormal data and the rest segments are normal.

The proposed method is implemented as follows. First, Algorithm 1 is applied to the time series of $X_1(n)$ and $X_2(n)$ for determination of KTPs with the minimum time interval $\delta = 5$.The algorithm stops at $K_{\max} = 114$ segments to meet with the requirement on δ. Fig. 3 presents the fitting error $E(K)$ as a function of K, where the variations of $E(K)$ become insignificant after the first five points. Thus, the data samples are divided into $K_{opt} = 6$ segments by 7 KTPs with the sampling indices $Q_{K_{opt}} = [1, 45, 155, 302, 353, 559, 700]$, which are also represented by the dash lines in Fig. 2.

Second, the values of the correlation directions of $\text{sign}_s(X_1, X_2)$ for the six segments are calculated to be -0.9606, 0.9779, 0.9322,

-0.9316, 0.0232 and 0.9244, and the corresponding hypothesis tests in Section 4.3 are implemented with $\alpha = 0.05$ with the results of H_0 is rejected for five segments except the 5th one; thus, the calculated correlation directions in the six segments are the same as the actual values $-1, 1, 1, -1, 0, 1$. Based on the calculated correlation directions, we detect three segments with $n \in [1, 45)$, $n \in [302, 353)$ and $n \in [353, 559)$ as the abnormal segments and the rest segments as the normal ones.

Third, the performance of the proposed method can be assessed by misclassification rate (MR),

$$\mathrm{MR} = \frac{N_m}{N}, \tag{8}$$

where N_m is the number of the samples being misclassified and N is the data length of the historical data set $\{X_1(n), X_2(n)\}_{n=1}^{N}$. By comparing with the actual normal and abnormal data points, it is revealed that $N_m = 13$ out of $N = 700$ points are misclassified, i.e., $\mathrm{MR} = 1.86\%$. Thus, it can be concluded that the proposed method yields a satisfactory result for the detection of normal and abnormal data segments.

Next, if a time delay is present, e.g., $X_{1,\tau}(n) := X_1(n + \tau)$ with

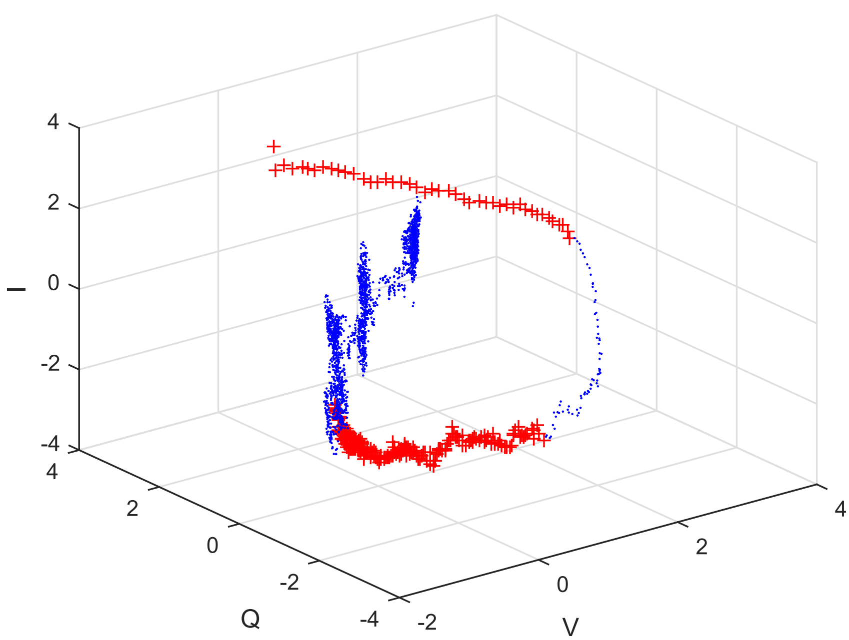

Fig. 10. The detected normal (dot) and abnormal (star) data segments for Example 2.

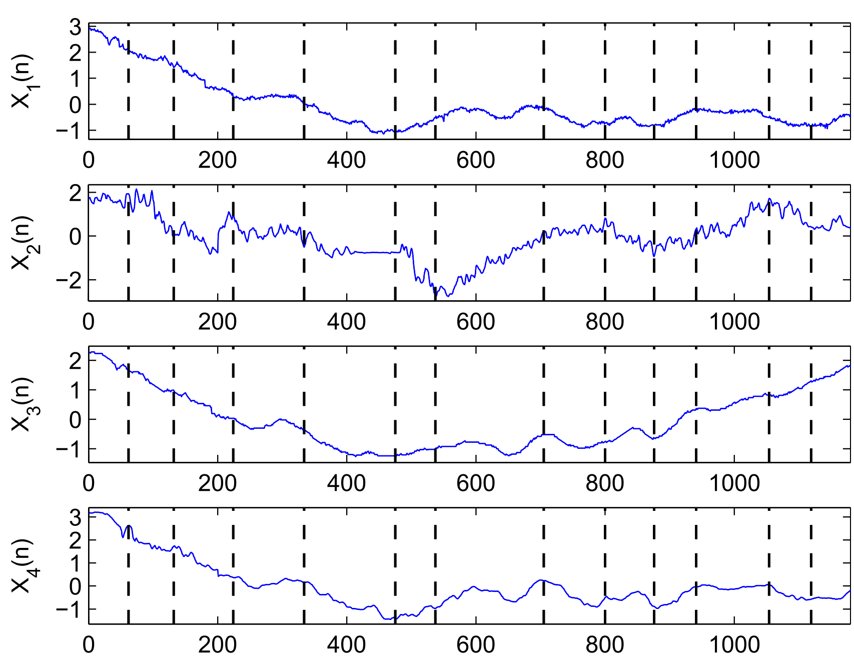

Fig. 11. The time sequence plots of normalized process variables with the dash vertical lines to separate segments for Example 3.

$\tau = 50$, the proposed method yields six segments for $\{X_{1,\tau}(n), X_2(n)\}_{n=1}^{650}$ with the sampling indices $Q_{K_{opt}} = [1, 49, 154, 312, 369, 515, 650]$, which are represented by the dash lines in Fig. 4. Clearly, the time delay causes a biased segmentation, e.g., $X_{1,\tau}(n)$ in the 2nd segment for $n \in [49, 154]$ changes its direction from being increasing to decreasing while $X_2(n)$ keeps increasing. In order to alleviate the negative effect of time delays, Algorithm 3 is applied with the obtained results of $E(K_{opt}(0, d))$ for $X_{1,\tau}(n)$ and $X_{2,d}(n) := X_2(n + d)$ given in Fig. 5. The optimal time delay estimate is $d_{opt} = 50$. The proposed method yields a satisfactory result for $X_{1,\tau}(n)$ and $X_{2,d_{opt}}(n)$ as shown in Fig. 6.

Example 2. Consider a feed water pump and its booster electrical pump from a large-scale thermal power generation unit at Shandong Province, China. According to physical principles of the two pumps, in the normal operating conditions, the outlet flow rate Q always keeps a positive correlation with both the pump rotating speed V and the electrical current I of the booster pump, that is, the three process variables are increasing or decreasing synchronously. Based on this fact, the correlation directions in normal conditions

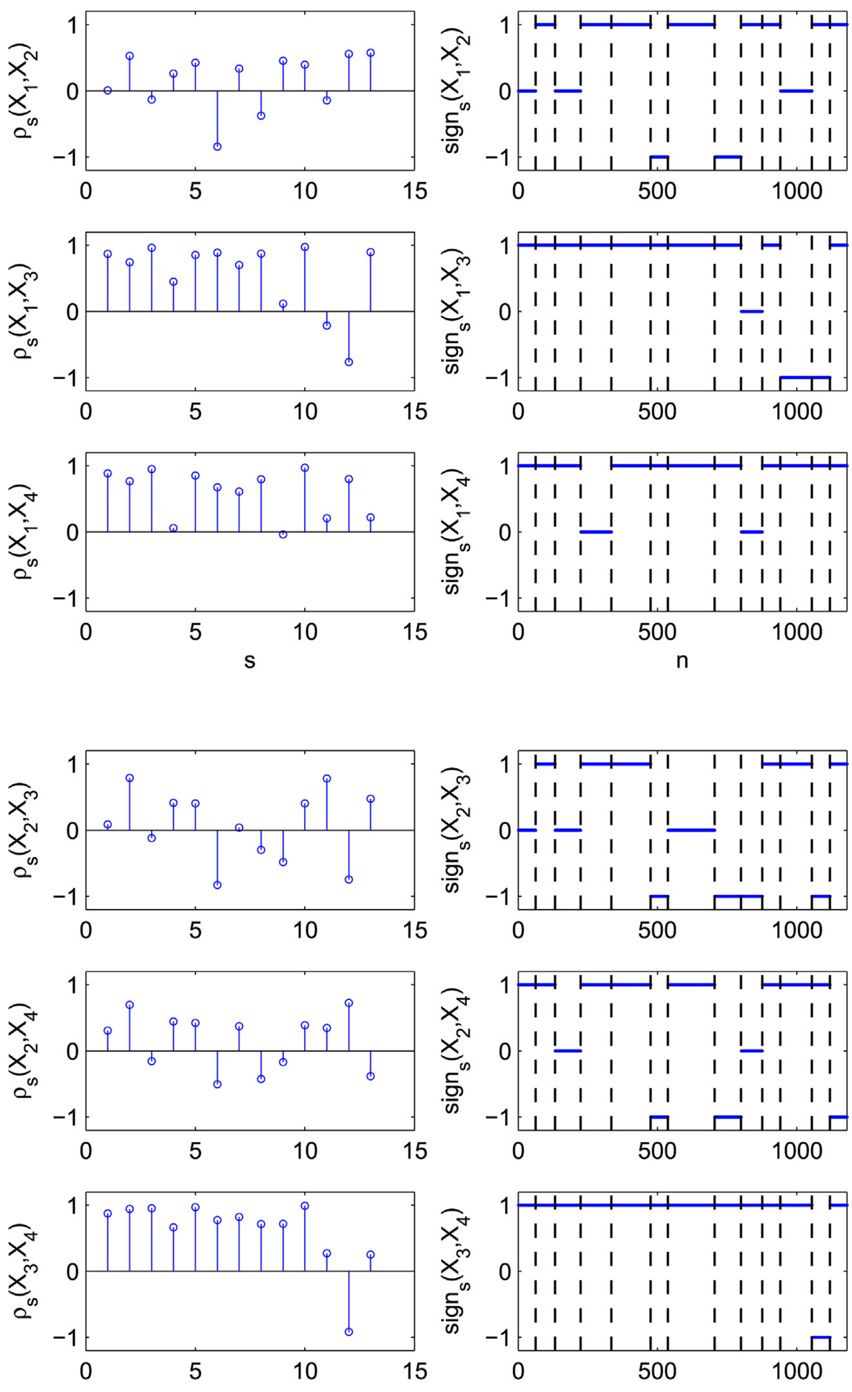

Fig. 12. The calculated correlation coefficients (left) and correlation directions (right) with the dash vertical lines to separate segments for Example 3.

are pre-defined as $\text{sign}(X_1, X_2) = 1, \text{sign}(X_1, X_3) = 1$ and $\text{sign}(X_2, X_3) = 1$. Fig. 7 presents the time sequence plots of the normalized time series $T = \{Q(n), V(n), I(n)\}$ for $n = 1, \ldots, 2240$ collected with the sampling period of 1 s.

The proposed method is implemented as follows. First, Algorithm 1 is applied to stop at $K_{\max} = 115$ confined by $\delta = 15$. Algorithms 2 and 3 tell that no time delays are present, and the number of segments is selected as $K_{opt} = 11$ shown as the star in the scatter plot between $E(K)$ and K in Fig. 8. The dash lines in Fig. 7 represent the time instants of KTPs with $Q_{K_{opt}} = [1, 356, 622, 687, 1080, 1300, 1597, 1769, 2006, 2125, 2195, 2240]$.

Second, based on the selected KTPs, the correlation analysis is carried out for each segment with $\alpha = 0.05$. The left three subplots of Fig. 9 show Spearman's rank correlation coefficients between the three pairs of process variables Q, V and I in each segment. The corresponding values of the correlation directions are given in the right subplots of Fig. 9. It is clearly observed that the pairs of (V, Q) and (V, I) always keep positive correlation till the 8-th segment, and experience some significant variations afterwards. As a result, the data points located in the 9-th segment with $n \in [2006, 2125]$ and the 11-th segment with $n \in [2195, 2240]$ are classified as abnormal data segments and the rests are normal.

The detected abnormal data segments are confirmed by the scatter plot in Fig. 10. The feedwater pump supplied water to the boiler drum that experienced an abnormal high drum pressure owing to an unbalance between the feed-coal flow and the generating power. The high drum pressure introduced a large resistance to the feedwater pump; as a result, the pump rotating speed V was increasing but the outlet flow rate Q kept decreasing. Thus, the feedwater pump was subject to an abnormal operating condition. By contrast, the relation between the outlet flow rate Q and the electrical current I of the booster pump was changing in the same direction, saying the abnormality in the feedwater pump was not prorogated to the booster pump.

Example 3. The large-scale coal-fired power generation unit in Example 2 is a complex industrial facility involving a large amount of process variables being monitored. The main operation of the power generation unit is to convert the heat energy of combustion by burning coal into the thermal energy of high-pressure high-temperature steam, and to generate the electricity via an electrical generator from the mechanical energy provided by a steam turbine. The operating status of the power generation unit as a whole can mainly be described by several process variables, including the generated power, feed-coal flow, air flow, furnace pressure, main steam pressure, main steam temperature, main steam flow, feed-water flow, drum level and condenser vacuum. Here we apply the proposed method to four selected process variables, namely, the generating power X_1, feed-coal flow X_2, main steam pressure X_3, and main steam flow X_4, in order to detect abnormal data segments from historical data sets. Note that the proposed method is applicable to all the above-mentioned process variables; however, the results for all these variables would be too cumbersome to present so that the four variables are selected to ease the presentation. Based on the physical laws among the four variables, when the power generation unit is in the normal condition, the four variables increase, decrease or keep constant in a synchronized manner. Thus, if $\text{sign}_s(X_i, X_j)$ takes the value -1 for $i \neq j (i, j \in [1, 4])$, then the segment s is classified as the abnormal one.

Some historical data points of $X_1(n), X_2(n), X_3(n)$ and $X_4(n)$ for $n = 1, \ldots, 1200$ are shown in Fig. 11. The proposed method is implemented as follows. First, using the minimum time interval $\delta = 60$, Algorithms 1, 2 and 3 are applied to estimate the time delays as $d_{1,opt} = 20$, $d_{2,opt} = 0$, $d_{3,opt} = 10$ and $d_{4,opt} = 20$, and to obtain the optimal number of segments as $K_{opt} = 13$. The estimated time delays are in line with physical relations among the four variables, e.g., a variation of X_2 as the energy source is expected to introduce a corresponding variation in X_1 and X_3 after a time period taken by the energy conversion processes. Thus, the historical data points are divided into 13 segments with the KTPs with $Q_{K_{opt}} = [1, 62, 132, 224, 334, 475, 537, 705, 800, 876, 941, 1054, 1119, 1180]$, which are marked by the vertical dash lines in

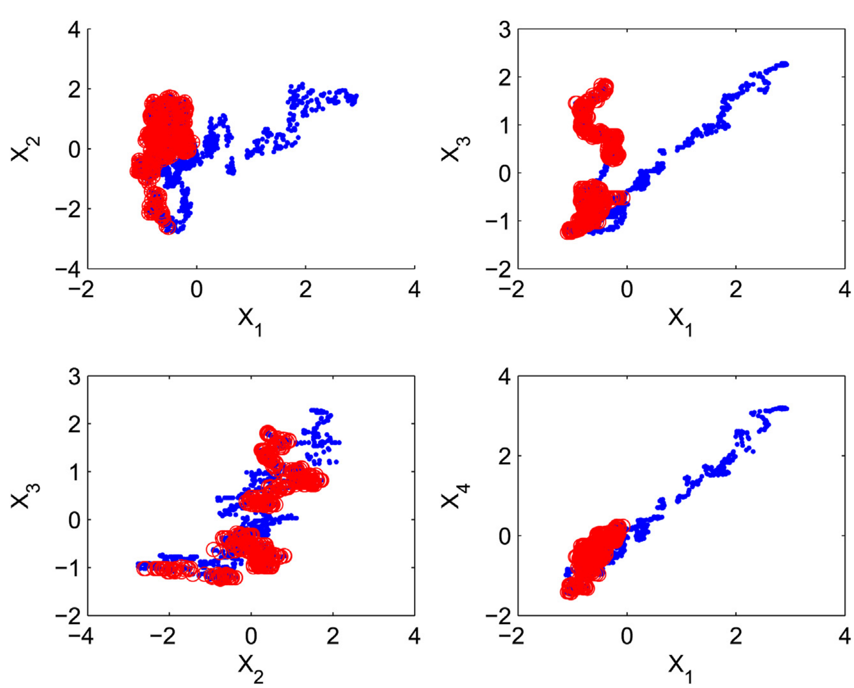

Fig. 13. The selected scatter plots of X_1, X_2, X_3 and X_4 for Example 3, where the detected abnormal segments are in red circles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 11. Second, based on the selected KTPs, the correlation directions are calculated in each segment with $\alpha = 0.05$. The left six subplots of Fig. 12 show Spearmans rank correlation coefficients between the paired process variables in each segment. The corresponding values of correlation directions are given in the right subplots of Fig. 12. Based on the calculated correlation directions, it is observed that the 6-th, 8-th, 9-th, 11-th, 12-th and 13-th segments are classified as the abnormal ones, and the rest segments are normal.

The detected abnormal segments are confirmed from selected scatter plots in Fig. 13, where the abnormal data segments are presented as the red circles. For the abnormal data segments in Fig. 13, both the feed-coal flow X_2 and the main steam pressure X_3 increase, while the generated power X_1 does not increase. This is inconsistent with the law of energy conservation. Thus, the power generation unit was subject to an abnormal condition. This abnormal condition indicates a low operation efficiency and deserves attentions from power plant operators for improvement in order to avoid occurrences of similar abnormalities in future.

6. Conclusion

In this paper, a method was proposed to detect abnormal data segments from historical data samples of multivariate systems. It is based on an observed fact that related process variables increase or decrease in a synchronized manner in normal conditions, while such correlation directions no longer hold for abnormal conditions. Key turning points (KTPs) were determined by finding the piecewise linear representation of multivariate time series, to be associated with the time instants of changing correlation directions. Spearman's rank correlation coefficients were calculated for each data segment between adjacent KTPs, and correlation directions were obtained by performing associated hypothesis tests on the correlation coefficients. Data segments were classified into normal or abnormal ones by comparing the calculated correlation directions with their counterparts in normal conditions obtained from process knowledge. Numerical and industrial examples illustrated the effectiveness of the proposed method.

References

Alrowaie, F.; Gopaluni, R.B.; Kwok, K.E. 2014. Alarm design for nonlinear stochastic systems. In: the 11th world congress on intelligent control and automation, Shenyang, China, pp. 473–479.

Beebe, D., Ferrer, S., Logerot, D., 2013. The connection of peak alarm rates to plant incidents and what you can do to minimize. Process Saf. Prog. 32, 72–77.

Björklund, S., 2003. A Survey and Comparison of Time-delay Estimation Methods in Linear Systems. Licentiate thesis no. 1061 SE-581 83. Linköping University, Linköping, Sweden.

Bransby, M.L., Jenkinson, J., 1998. The Management of Alarm Systems. Health and Safety Executive.

Brooks, R., Thorpe, R., Wilson, J., 2004. A new method for defining and managing process alarms and for correcting process operation when an alarm occurs. J. Hazard. Mater. 115, 169–174.

Charbonnier, S., Garcia-Beltan, C., Cadet, C., Gentil, S., 2005. Trends extraction and analysis for complex system monitoring and decision support. Eng. Appl. Artif. Intell. 18, 21–36.

Cheung, J.Y., Stephanopoulos, G., 1990. Representation of process trends: Part I. A formal representation framework. Comput. Chem. Eng. 14, 495–510.

Chung, F.L., Fu, T.C., Ng, V., Luk, R.W.P., 2004. An evolutionary approach to pattern-based time series segmentation. IEEE Trans. Evol. Comput. 8, 471–489.

Duncan, S.R.; Bryant, G.F. 1996. A new algorithm for segmenting data from time series. In: IEEE Conference on Decision & Control, Kobe, Japan. pp. 3123–3128.

Fu, T.C., Chung, F.L., Luk, R., Ng, C.M., 2007. Stock time series pattern matching: template-based vs. rule-based approaches. Eng. Appl. Artif. Intell. 20, 347–364.

Gupta, A., Giridhar, A., Venkatasubramanian, V., Reklaitis, G.V., 2013. Intelligent alarm management applied to continuous pharmaceutical tablet manufacturing: an integrated approach. Ind. Eng. Chem. Res. 52, 12357–12368.

Han, L., Gao, H., Xu, Y., Zhu, Q., 2016. Combining fap, map and correlation analysis for multivariate alarm thresholds optimization in industrial process. J. Loss Prev. Process Ind. 40, 471–478.

Janusz, M.E., Venkatasubramanian, V., 1991. Automatic generation of qualitative description of process trends for fault detection and diagnosis. Eng. Appl. Artif. Intell. 4, 329–339.

Keogh, E., Chu, S., Hart, D., Pazzani, M., 2004. Segmenting Time Series: a Survey and Novel Approach. Data Mining in Time Series Databases. World Scientific, pp. 1–21.

Kim, I., 1994. Computerized systems for online management of failures: a state-of-the-art discussion of alarm systems and diagnostic systems applied in the nuclear industry. Reliab. Eng. Syst. Saf. 44, 279–295.

Macdonald, D., 2004. Practical Hazops, Trips and Alarms. Elseviers.

Pariyani, A., Seider, W.D., Oktem, U.G., Soroush, M., 2010. Incidents investigation and dynamic analysis of large alarm databases in chemical plants: a fluidized-catalytic-cracking unit case study. Ind. Eng. Chem. Res. 49, 8062–8079.

Perng, C.S.; Wang, H.; Zhang, S.R.; Parker, D.S. 2000. Landmarks: a new model for similarity-based pattern querying in time series databases. In: Proceedings of the 16th International Conference on Data Engineering, San Diego, California, pp. 33–42.

Pratt, K.B., Fink, E., 2002. Search for patterns in compressed time series. Int. J. Image Graph. 2, 89–106.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical Recipes: the Art of Scientific Computing, third ed. Cambridge University Press.

Schneider, P., Eberly, D.H., 2002. Geometric Tools for Computer Graphics. Morgan Kaufmann.

Soderstrom, T., Stoica, P., 1989. System Identification. Prentice Hall.

Stauffer, T., Clarke, P., 2016. Using alarms as a layer of protection. Process Saf. Prog. 35, 76–83.

Wang, J., Yang, F., Chen, T., Shah, S.L., 2016. An overview of industrial alarm systems: main causes for alarm overloading, research status, and open problems. IEEE Trans. Autom. Sci. Eng. 13, 1045–1061.

Zang, H., Li, H., 2014. Optimization of process alarm thresholds: a multidimensional kernel density estimation approach. Process Saf. Prog. 33, 292–298.

Zhou, B., Ye, H., 2016. A study of polynomial fit-based methods for qualitative trend analysis. J. Process Contr. 37, 21–33.

Zhu, J., Shu, Y., Zhao, J., Yang, F., 2014. A dynamic alarm management strategy for chemical process transitions. J. Loss Prev. Process Ind. 30, 207–218.