Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 6, No. 1, 2023

Fault Detection Based on Complete Information Principal Component Analysis for Electric Submersible Pump

Jie Yuan*, Yinchang Du, Jiankui Xu, Jianshen Liu

Designing Institute of Engineering, Offshore Oil Engineering Co.,Ltd, Tianjin, China

* **Corresponding author**: yuanjie7@cnooc.com.cn

Abstract: Electric submersible pump (ESP) is the key production equipment for the offshore oil industry. As a result of poor working condition, fault and failures happen frequently and affect the production effectiveness. It is significant to detect the fault and failure in time. The fault detection for ESP mainly depends on the expert experience for decades. The data-driven fault detection methods are widely utilized in recent years with the rapid development of sensor technology, where the PCA is the most widely used methods. However, the data collected from the front line faces the imprecision problem, which reduces the accuracy of data-driven fault detection methods. The interval arithmetic is an effective way to handle imprecision data. However, traditional PCA and interval PCA algorithms cannot handle nonlinear problem. Aiming at this problem, the complete information principal component analysis methods are adopted for the fault detection of electric submersible pump in this paper. Compared with PCA and CPCA methods, the CIPCA method has better detection robustness when more inaccurate point value data.

Keywords: Interval algorithm, Principal component analysis, Electric submersible pump, Fault detection.

1. Introduction

As the key equipment of offshore oil industry, electric submersible pump (ESP) transport quantities of oil from underground to the surface. However, ESP works under high temperature and high pressure conditions which leading to failures frequently. Severe economic losses will happen once the electric submersible pump failure occurs. Efficient fault detection of electric submersible pump can help production workers to judge faults quickly and accurately, reduce the equipment maintenance time, and help producers to find early failures of electric submersible pump in time. Therefore, it is of great significance to study the fault detection of electric submersible pump(1).

Traditional fault detection of electric submersible pumps in offshore oil and gas wells mainly relies on manual calculation and expert experience(2). With the rapid development of computing technologies and sensors, data-driven methods have become effective approaches for addressing the fault detection problem and have become increasingly popular in recent years. Zhao et al(3). proposed a full-condition monitoring method with cointegration and slow feature analysis for nonstationary dynamic chemical processes. It is not necessary to extract the process knowledge for these methods because all the information is stored in the process data. The reliability of the fault detection model is strongly dependent upon the quality of the available data. Fault detection methods such as principal component analysis (PCA) (4)and kernel PCA (KPCA)(5) assume that the process data are precisely known and point-valued. However, in practice, the data are typically contaminated with uncertainties due to various factors, which include measurement noise, sensor degradation from age, and severe working environments (6). Imprecise measurements with low reliability are prominent and may deteriorate the identification performance of the data-driven methods that are discussed above. The development of fault detection methods for the process industry with imprecise data is an urgent problem.

A substantial amount of work has been done by researchers on the processing of imprecise data. However, classic data analysis cannot satisfactorily deal with data with variability and uncertainty. Symbolic data analysis (SDA) introduced several representations for data types with data variability, among which the interval representation is typical. Intervals naturally arise from the description of ranges of values, such as daily temperature and daily stock price variations. Numerical examples were provided by Lev V. to demonstrate that imprecise interval information may improve the robustness performance of the SVM algorithm(7). A new clustering algorithm that is based on the learning of a self-organizing map for interval data is proposed in the literature, in which the number of clusters is determined automatically and no a priori hypothesis is required. Petridis et al. used interval numbers to predict annual sugar production based on populations of measurements and to improve the prediction accuracy(8). Several artificial neural networks were extended to interval fields and have exhibited satisfactory performance in industry applications with imprecise measurement data. In addition, The interval PCA approaches were introduced by Cazes et al. and Chouakria et al. known as the vertices PCA (VPCA) and the centers PCA (CPCA), respectively. The interval PCA methods are employed for process monitoring and yielded more robust results(9).

In this paper, aiming at improving effectiveness of PCA method when facing the imprecision measurement problem, several interval PCA methods are applied for the fault detection of electric submersible pump. Sections 2 is a brief introduction of interval PCA. Section 3 discusses the simulation of interval PCA to electric submersible pump FDI. Section 4 is the conclusion that the IPCA performs better than the point valued PCA when the collected data is imprecise.

2. Preliminary

2.1. PCA

In recent years, principal components analysis(10) is widely used to extract the main relationship of variables in a multidimensional data set. It reduces the complexity of the multivariate data via a projection into a subspace which preserves maximum variance of the original space in less number of dimensions. Let us consider $X \in R^{n\times m}$ the data matrix containing *n* samples of *m* process variables. Before applling PCA, matrix *X* must be normalized to zero mean and unit variance. Then, the PCA linear projection can be expressed as

$$T = XP,\ X = TP^T \tag{1}$$

where $T \in R^{n\times m}$ is the principal component matrix and $P \in R^{m\times m}$ contains the principal vectors which are the eigenvectors associated with the eigenvalues of the covariance matrix $\sum$ of X.

$$\begin{aligned}\sum &= \frac{1}{N-1}X^T X \\ &= P\Lambda P^T = \begin{bmatrix} P_l & P_{m-l}\end{bmatrix}\begin{bmatrix}\Lambda_l & 0 \\ 0 & \Lambda_{m-l}\end{bmatrix}\begin{bmatrix} P_l^T \\ P_{m-l}^T\end{bmatrix}\end{aligned} \tag{2}$$

where $PP^T = P^T P = I_m$, and Λ is a diagonal matrix that contains in its diagonal the eigenvalues of $\sum$ sorted in decreasing order. Then, the most important step in performing PCA modeling is the determination of the number of components to be retained. Several ways can be applied to determine the number of main components, in this paper, the percentage of accumulated contribution of variances can be used as the parameter.

The transformation matrix $P \in R^{m\times l}$ is generated by choosing l eigenvectors of P corresponding to largest l eigenvalues. The reduced dimension space X can be denoted as

$$X = XP_l P_l^T \tag{3}$$

The residual space, also denoted E, is the residual between X and X.

$$E = X - X \tag{4}$$

At each sampling time k, the measurement vector is $x(k) = \left[x_1(k)\ x_2(k)\ldots x_m(k)\right]$, and its estimation is $x(k) = \left[x_1(k)\ x_2(k)\ldots x_m(k)\right]$,and estimation error can be expressed as:

$$e(k) = x(k) - x(k) \tag{5}$$

When it comes to the fault detection, the established PCA model describes the relationship between normal process data in principal and residual spaces. Fault events will be detected by referencing the observed behavior against this model especially in residuals which are expected to be null in the fault-free case, and different from zero in the presence of faults.

2.2. IPCA

In this section, several interval PCA methods are introduced. As the interval PCA approach is an extended version of the classical PCA approach applied to interval data. the interval data are introduced first.

2.2.1 interval data

Interval arithmetic is the foundation of many interval algorithms(11). Interval numbers are the basis of arithmetic operations and are defined as $X = \left[\underline{x}, \overline{x}\right] \in I(R), \underline{x} \le \overline{x}$, where $\underline{x}$ and $\overline{x}$ are real numbers and X is an interval number. If the width of X is denoted as $w(X)$, then $w(X) = \overline{x} - \underline{x}$. The median of X can be denoted as $m(X) = (\underline{x} + \overline{x})/2$.

If $X \in R^{n\times m}$ is the conventional data matrix containing n samples of m process variables in normal condition, where $x_j(k)$ is the k[th] observation of the j[th] variable. All the observation data is considered imprecise and expressed in interval format, then the new interval matrix [*X*] is formalized as a set of interval valued observation as below

$$[X] = \begin{bmatrix} \left[\underline{x_1}(1),\ \overline{x_1}(1)\right] & \cdots & \left[\underline{x_m}(1),\ \overline{x_m}(1)\right] \\ \vdots & \cdots & \vdots \\ \left[\underline{x_1}(n),\ \overline{x_1}(n)\right] & \cdots & \left[\underline{x_m}(n),\ \overline{x_m}(n)\right]\end{bmatrix} \tag{6}$$

When defining an interval model of the measured data, the radius of the interval observation is given by the measurement error $x_j^r(k)$, while the center of the interval is given by the measurement $x_j^c(k)$. Thus, yielding the following interval construction formulas, representing respectively, standard form and midpoint-radius form.

$$\left[x_j(k)\right] = \left[x_j^c(k) - x_j^r(k), x_j^c(k) + x_j^r(k)\right] \tag{7}$$

where $x_j^c(k) = \dfrac{\left(\overline{x_j}(k) + \underline{x_j}(k)\right)}{2}$ and $x_j^r(k) = \dfrac{\left(\overline{x_j}(k) - \underline{x_j}(k)\right)}{2}$.

2.2.2. *Centers* PCA

Most interval PCA methods are also based on the analysis of transformed matrices from initial interval data as vertices of hyper-cubes or as centers and radii. CPCA(12) decomposes the correlation matrix of the centers coded matrix X^C and it projects the vertices as supplementary points in the factorial subspace.

$$X^C = \begin{pmatrix} x_1^c(1) \cdots x_m^c(1) \\ \vdots \quad \ddots \quad \vdots \\ x_1^c(n) \cdots x_m^c(n) \end{pmatrix} \tag{8}$$

where $x_j^c(k)$ is the midpoint or center of the interval at hand. The CPCA algorithm for interval data can be resumed in the following steps:

• Step 1: Calculate the centers coded matrix X^C as in (16).

• Step 2: Apply a classical PCA on the centers matrix, based on the symbolic covariance given in (9) and (10), with $T_1, \ldots, T_m$ being the principal components of this PCA, and $p_1, \ldots, p_m$ their corresponding eigenvectors.

• Step 3: Determine the new interval principal components as:

$$\begin{cases} \underline{t}_j(k) = \sum_{i=1, p_{ij}<0}^{m} \bar{x}_i(k) p_{ij} + \sum_{i=1, p_{ij}>0}^{m} \underline{x}_i(k) p_{ij} \\ \bar{t}_j(k) = \sum_{i=1, p_{ij}<0}^{m} \underline{x}_i(k) p_{ij} + \sum_{i=1, p_{ij}>0}^{m} \bar{x}_i(k) p_{ij} \end{cases} \tag{9}$$

where p_{ij} is the i[th] element of the j[th] column of eigenvector matrix P.

• Step 4: Compute interval estimates based on CPCA model for the first l components as:

$$\begin{cases} \underline{x}_j(k) = \sum_{i=1, c_{ij}<0}^{m} \bar{x}_i(k) c_{ij} + \sum_{i=1, c_{ij}>0}^{m} \underline{x}_i(k) c_{ij} \\ \bar{x}_j(k) = \sum_{i=1, c_{ij}<0}^{m} \underline{x}_i(k) c_{ij} + \sum_{i=1, c_{ij}>0}^{m} \bar{x}_i(k) c_{ij} \end{cases} \tag{10}$$

where c_{ij} is the i[th] element of the j[th] column of matrix C.

In a word, CPCA performed on interval input data is based on a numerical centers codification of the data, a treatment with classical PCA analysis technique, and finally a transformation of classical results into interval description.

3. Methodology

3.1. Complete information PCA

In order to seize more information within interval observations, Huiwen et al. proposed a new interval PCA method called complete information based PCA (CIPCA). By defining the inner product of hyper-cubes divided into informative grid data, and based on a rather analytic approach, CIPCA accomplishes the derivation of interval-valued principal components and transforms PCA modeling into the computation of some inner products. Thus, leading to more accurate results and providing an efficient and effective way for conducting PCA on large-scaled interval data.

According to CIPCA, given any two interval-valued variables $\left[X_j\right]$ and $\left[X_{j'}\right]$, the inner product is defined as

$$\left\langle \left[X_j\right], \left[X_{j'}\right] \right\rangle = \frac{1}{4}\left(\underline{x}_j(k) + \bar{x}_j(k)\right)\left(\underline{x}_{j'}(k) + \bar{x}_{j'}(k)\right) \tag{11}$$

Based on the above definitions of interval norm and inner product, and with all data units pre-processed, the covariance matrix of $X_{n\times m}$ is given by

$$\Sigma = \frac{1}{n}\begin{pmatrix} \langle [X_1],[X_1] \rangle & \langle [X_1],[X_2] \rangle & \cdots & \langle [X_1],[X_m] \rangle \\ \langle [X_2],[X_1] \rangle & \langle [X_2],[X_2] \rangle & \cdots & \langle [X_2],[X_m] \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle [X_m],[X_1] \rangle & \langle [X_m],[X_2] \rangle & \cdots & \langle [X_m],[X_m] \rangle \end{pmatrix} \tag{12}$$

The CIPCA method for interval-valued data can be summarized in the following steps:

• Step 1: Compute the covariance matrix Σ of the interval data matrix $\left[X\right]$ using (12).

• Step 2: Perform an eigen-decomposition of the covariance matrix Σ, where $\lambda_1, \lambda_2, \cdots, \lambda_m$ and $p_1, p_2, \cdots, p_m$ are the resulting eigenvalues and eigenvectors respectively.

• Step 3: Calculate the interval principal components using the following formulas

$$\begin{cases} \underline{t}_j(k) = \sum_{i=1}^{m} p_{ij}\left(\tau \underline{x}_i(k) + (1-\tau)\bar{x}_i(k)\right) \\ \bar{t}_j(k) = \sum_{i=1}^{m} p_{ij}\left((1-\tau)\underline{x}_i(k) + \tau \bar{x}_i(k)\right) \end{cases} \tag{13}$$

with

$$\tau = \begin{cases} 0, & p_{ij} \le 0 \\ 1, & p_{ij} \ge 0 \end{cases}$$

Step 4: Compute the interval estimates from CIPCA model as:

$$\begin{cases} \underline{x}_j(k) = \sum_{q=1}^{m} C_{lqj}\left(\tau \underline{x}_q(k) + (1-\tau)\bar{x}_q(k)\right) \\ \bar{x}_j(k) = \sum_{q=1}^{m} C_{lqj}\left((1-\tau)\underline{x}_q(k) + \tau \bar{x}_q(k)\right) \end{cases} \tag{14}$$

with the same condition on τ, and given that $C_l = P_l P_l^T$.

Based on the interval arithmetic, the estimated error between the estimated values and the process data,

$$\begin{cases} \underline{e}_{ij} = \underline{x}_{ij} - \hat{\underline{x}}_{ij} \\ \bar{e}_{ij} = \bar{x}_{ij} - \hat{\bar{x}}_{ij} \end{cases} \tag{15}$$

3.2. Off-line model building based on CIPCA

The CIPCA method is to perform interval PCA on the principle space and the residual space. The establishment steps of modeling are as follow,

Step 1: Convert the collected imprecision data $\mathbf{x} \in \mathbb{R}^{n\times m}$ into interval form,

$$[x_{n\times m}] = [\underline{x}_{n\times m}, \bar{x}_{n\times m}] \tag{16}$$

Step 2: Standardize the interval value data,

$$\frac{[x_{n\times m}]-E([\mathbf{x}_j])}{\sqrt{D([\mathbf{x}_j])}}=\left[\frac{\underline{x}_{n\times m}-E([\mathbf{x}_j])}{\sqrt{D([\mathbf{x}_j])}},\frac{\overline{x}_{n\times m}-E([\mathbf{x}_j])}{\sqrt{D([\mathbf{x}_j])}}\right] \quad (17)$$

Step 3: perform CIPCA introduced in section A on the standardize matrix $[x_{n\times m}]$.

Step 4: Two monitoring statistics can be obtained, which are denoted as statistics and statistics respectively. The specific calculation formula is as follows

$$IT^2_i=\left\|\frac{[\mathbf{t}_{i,:}]}{\left[\mathbf{\Lambda}_{CI,l}\right]^{1/2}}\right\|^2 \quad (18)$$

$$ISPE_i=\left\|[e_{i,:}]\right\|^2=\sum_{j=1}^{m}\left\|\mathbf{e}_{ij}\right\|^2 \quad (19)$$

where $\Lambda=diag(\lambda_1,\lambda_2,\cdots\lambda_l)$ is the diagonal matrix with the principle components eigenvalues as diagonal elements.

Step 5: After all the off-line monitoring statistics of modeling samples are calculated, kernel density estimation method can be used to obtain the control limit.

3.3. On-line fault detection based on CIPCA

The online detection steps of CIPCA method can be expressed as,

Step 1: Convert the online testin imprecision data $\mathbf{x}_{new}\in\mathbb{R}^{1\times m}$ into interval form, $[x_{j,new}]=[\underline{x}_{j,new},\overline{x}_{j,new}]$.

Step 2: Standardize the interval value data with Equ(17).

Step 3: perform CIPCA introduced in section A on the standardize sample $[x_{j,new}]$.

Step 4: IT^2_i and $ISPE_i$ can be obtained based on Equations (18) and (19), respectively.

Step 5: Compare the statistics with control limit and get the fault detection conclusion.

The offline modeling and online fault detection can be expressed in Fig.1,

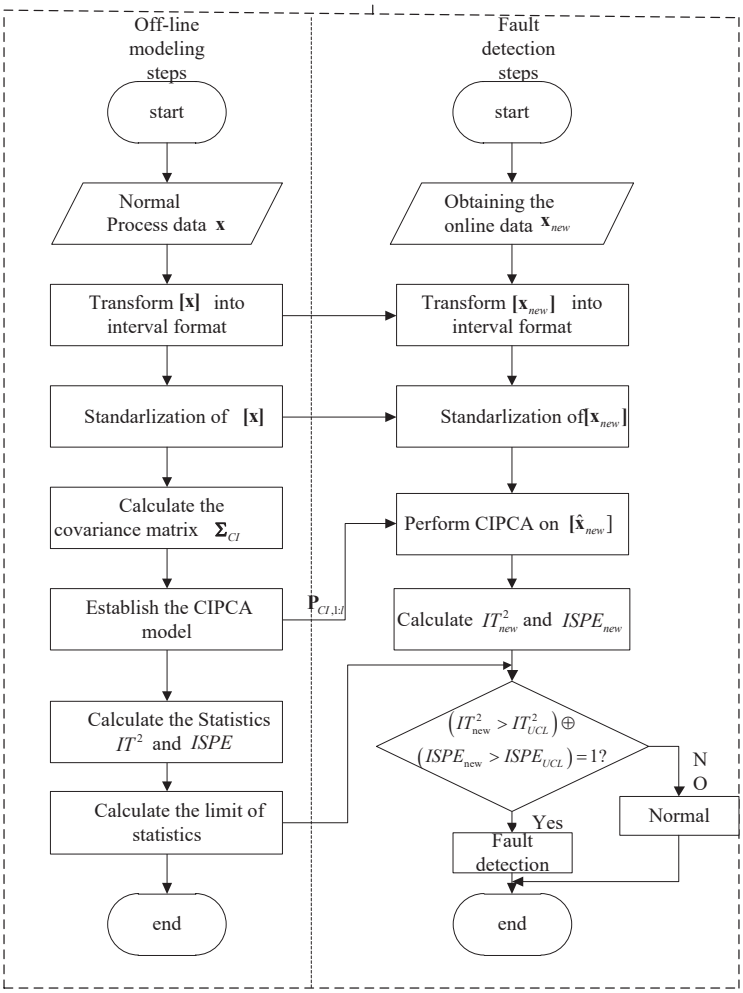

Figure 1. Flowchart of modeling and online fault detection

4. Simulation

In this section, a few simulations are performed to show the effectiveness of interval PCA when it faces the problem of data imprecision.

Firstly, 1000 pieces of production records are selected from history production database. The records are collected from 700 electric submersible pumps in offshore oil industry. Among the 1000 records, 800 are normal operation condition while other 200 are with different faults. All the records are of unequal length, high dimensional and low imprecision. Before the records are utilized for fault detection, necessary procedures are carried out to make each record equal length with 300 sample points. Twelve variables are selected as

character variables to identify the fault of ESP. The variables are listed in TABLE.1

Table 1. Selected Features variables of ESD

Variable	Unit	Variable	Unit
Pressure of Oil	Mpa,	Rotational Speed of ESD	r/min
Outlet Pressure of ESD	Mpa,	Acceleration of ESD	g
Intlet Pressure of ESD	Mpa,	Gas Oil Ratio	%
Fluid Pressure	Mpa,	Oil Gas Ratio	%
Current of ESD	A	Water Gas Ratio	%
Voltage of ESD	V	Oil Density	g/cm3

There are various of fault for the ESD. Despite the manual operation fault, the leaking of oil pipe, low load of the ESD, three-phase unbalanced of the ESD, over load of the ESD, insufficient of the water supply are the most common fault caused by unclear inner mechanic reason. The common faults are listed in TABLE.2

Table 2. Common faults of ESD

Index	Faults
1	Oil Pipe Leaking
2	Low Load of ESD
3	Three-phase Unbalanceof ESD
4	Over Load of ESD
5	Insufficient Water Supply

Firstly, three data-models are established. 100 samples are selected as the modeling data for PCA model, CPCA model, CIPCA model are checked by expert and are believed precise, relatively. The interval data for CPCA and CIPCA is transformed from the same model-building data as that in PCA model based on GMM algorithm with one time variance. While for the online cases, the test data contains more imprecision problem with 3 times variances.

Two fault cases are chosen from the fault records to verify the online fault detection ability. The Oil pipe leaking fault occurs in the first case from 201 sample point to the 300 sample point. The fault detection results of PCA, CPCA and CIPCA models are shown in Fig.2, Fig.3 and Fig.4

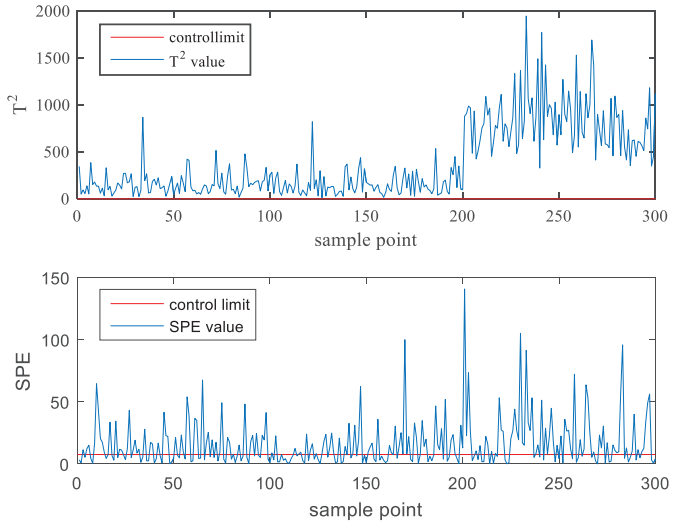

Figure 2. Fault detection results of PCA model for case1

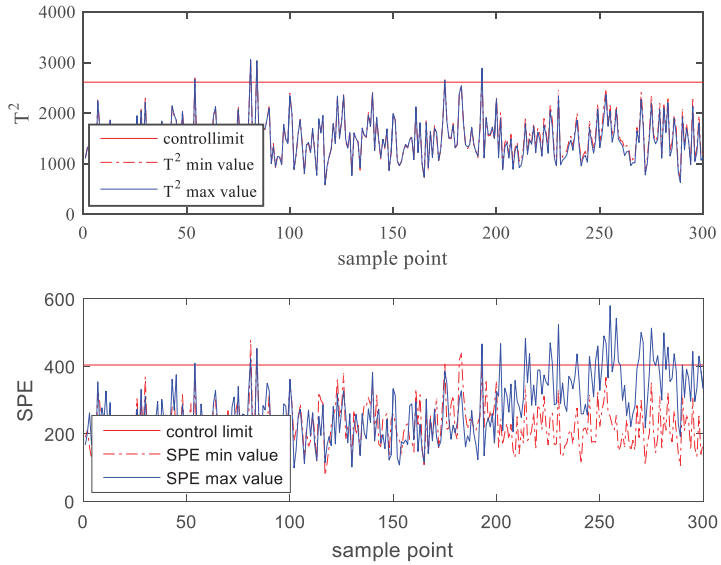

Figure 3. Fault detection results of CPCA model for case1

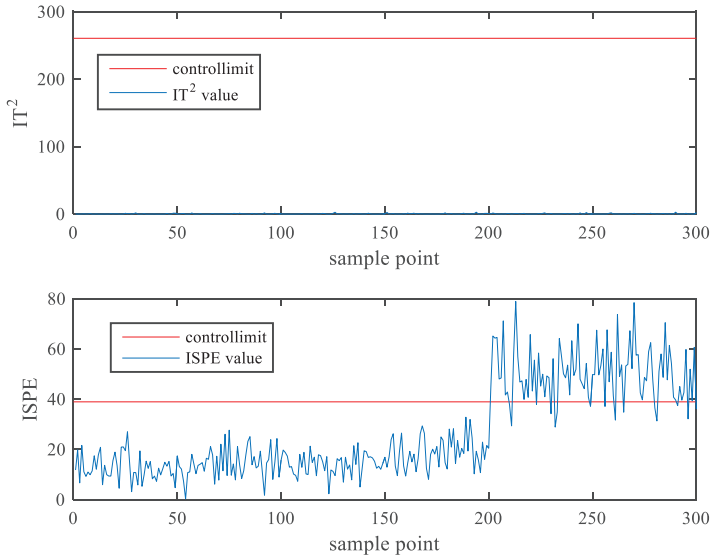

Figure 4. Fault detection results of CIPCA model for case1

As the test data contains much noise, the traditional PCA model regards the samples as fault sample. Therefore, nearly all the samples are tested fault sample as shown in Fig.2. For the interval model can handle the imprecision problem well in Fig.3. Almost all the normal samples are tested. However, the fault samples have low identification. The CIPCA utilized more data information and identify the fault in time and accurately as shown in Fig4.

In case 2, the three-phase unbalanced fault is introduced from 200 sample point to the end. The traditional PCA cannot separated the fault from the normal point. The CPCA can identify most fault sample point with a few false alarm. The CIPCA identify most the fault samples.

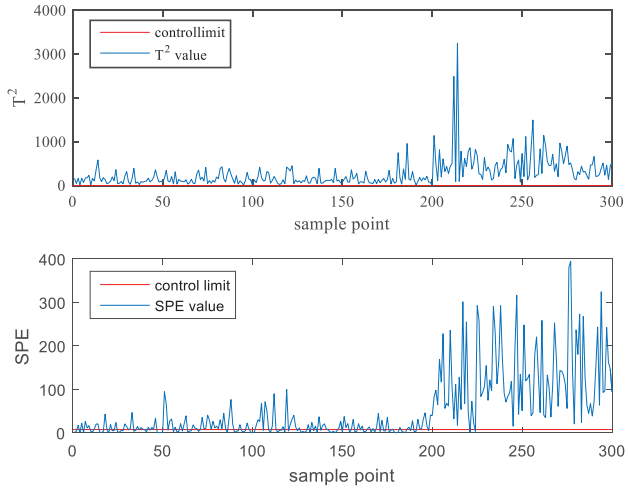

Figure 5. Fault detection results of PCA model for case2

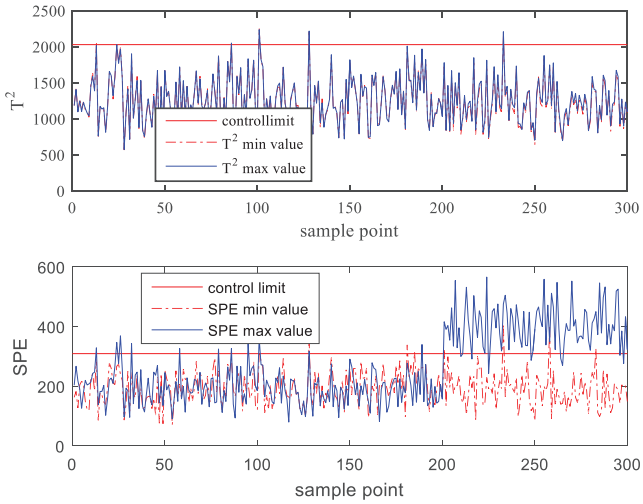

Figure 6. Fault detection results of CPCA model for case2

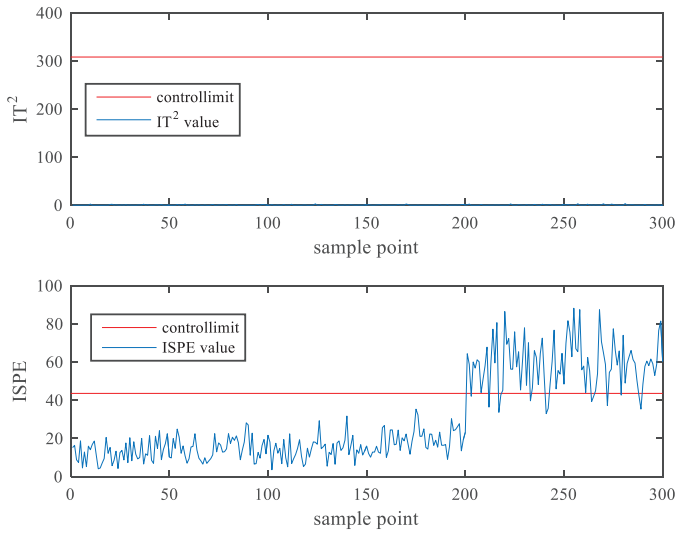

Figure 7. Fault detection results of CIPCA model for case2

5. Conclusion

Most of the data-driven fault detection methods of electric submersible pump are affected by the data imprecision. The interval data algorithms provide a way in handling this problem. The CPCA method can identify more fault and perform more sensitive than the point valued traditional PCA. However, the traditional and modified PCA algorithms mainly handle the linear relationship between variables. Some nonlinear relationship cannot be identified by the PCA algorithms. In this paper, the complete information PCA is applied to identify the faults and all the information between variables are considered and reserved. Better detection results are received.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No.61903272), the Natural Science Foundation of Tianjin (No.20JCQNJC01670) and Science and Technology on Space Intelligent Control Laboratory (No. HTKJ2019KL502008).

References

[1] Gupta S, Saputelli L, Nikolaou M, editors. Big Data Analytics Workflow to Safeguard ESP Operations in Real-Time. Spe North America Artificial Lift Conference & Exhibition; 2016.

[2] Zhang P, Chen T, Wang G, Peng C. Ocean Economy and Fault Detection of Electric Submersible Pump applied in Floating platform. International Journal of e-Navigation and Maritime Economy. 2017;6(C):37-43.

[3] Zhao C, Wang F, Gao F, Zhang Y. Enhanced Process Comprehension and Statistical Analysis for Slow-Varying Batch Processes. Industrial & Engineering Chemistry Research. 2008;47(24):9996-10008.

[4] Zhao C, Gao F, Wang F. Nonlinear Batch Process Monitoring Using Phase-Based Kernel-Independent Component Analysis−Principal Component Analysis (KICA−PCA). Industrial & Engineering Chemistry Research. 2009.

[5] Deng X, Deng J. Incipient Fault Detection for Chemical Processes Using Two-Dimensional Weighted SLKPCA. Industrial & Engineering Chemistry Research. 2019.

[6] Li G, Hu Y. An enhanced PCA-based chiller sensor fault detection method using ensemble empirical mode decomposition based denoising. Energy & Buildings. 2019;183(JAN.):311-24.

[7] Utkin LV. An imprecise extension of SVM-based machine learning models. Neurocomputing. 2019;331:18-32.

[8] Petridis V, Kaburlasos VG. FINkNN: A Fuzzy Interval Number k-Nearest Neighbor Classifier for Prediction of Sugar Production from Populations of Samples. Journal of Machine Learning Research. 2003;4(1):17-37.

[9] Ait-Izem T, Harkat MF, Djeghaba M, Kratz F. On the application of interval PCA to process monitoring: A robust strategy for sensor FDI with new efficient control statistics. Journal of Process Control. 2018;63:29-46.

[10] Lu C, Feng J, Chen Y, Liu W, Lin Z, Yan S. Tensor Robust Principal Component Analysis with A New Tensor Nuclear Norm. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020;42(4):925-38.

[11] Menini L, Possieri C, Tornambe A. Design of high-gain observers based on sampled measurements via the interval arithmetic. Automatica. 2021(131-).

[12] Das S, Bora PK, Gogoi AK, editors. Subtractive clustering of vertices for CPCA based animation geometry compression. Indian Conference on Computer Vision; 2010.