LUT Computer Vision and Pattern Recognition Laboratory

ADAML, BM20A6100

Supervision: MSc Ramona Maraia, Prof. Heikki Haario

TITLE OF THE REPORT:

Fault detection in an industrial process

Date :26-10-2020

Group: Fault Detection Group B

Yang Tuo (StuNo:0589715)

Subhashree Rautray (StuNo:0592964)

Contents

1. Background and Data

The assignment material consists of several cases of industrial process data, as produced by the Tennessee Eastman process simulator (see the provided article by Russel et al.). There are 22 different data sets, each consisting of 960 observations of 52 process variables. One of them, the file d00_te.dat contains observations from the process under normal conditions, while all the rest (d01_te.dat, ..., d21_te.dat) present a fault, some behavior that deviates from the reference case in d00 te.dat.

2. Problem Solutions

Datasets we have chosen: d00_te.dat, d04_te.dat, d11_te.dat, d15_te.dat

Because the d00_te.dat contains observations from the process under normal conditions, we plan to use it as training dataset to help us to decide how many variables we need to take to help calculate the T2 and Q statistics, and it will be also used as the parameters inside the T2 and Q's formulas to get these two statistics' values.

But the left three datasets: d04_te.dat, d11_te.dat, d15_te.dat will be used as testing datasets to calculate those two statistics' values mentioned above. Meanwhile, the T2 and Q's threshold values will be also gained by using these three datasets as parameters of their threshold formulas.

The whole algorithm for fault detection of industrial process data mainly adopt two methods: PCA and DPCA, and use T2 and Q statistics to detect faults, by the way, we added one new statistic combining those two we mentioned above for detecting faults. From figures drawn out we could judge where the fault begins and where it gets back to the normal, but for knowing which variable has caused the fault, we draw the contributions of fault plots to more vividly identify problemed variable.

The whole processes of fault detection for one dataset are as follows (let's take d04_te as the example):

Step 1: load normal data and fault data (d00_te.dat and d04_te.dat) from data storage files, then we get two variables d00_te and d04_te, respectively doing scaling (standardization or centralization) operations for them, the matlab codes are as follows:

```
fault_data = fault_dataset{k,2};
scaled_fault_data = scale(fault_data);   %standardise the data
normal_data = d00_te;
scaled_normal_data = scale(d00_te);
```

Figure 2.1 scaling codes for dataset

Step 2: for identifying the number of principal components we need to choose, doing svd decomposition for covariance matrix of fault data(d04_te), and we will get results of eigen values ranked in the diagonal matrix S in the descending number like $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. and we set cumulative contribution rate as 85%, use the formula

sum($\lambda_1+\lambda_2+\ldots+\lambda_n$)/sum(diag(S)) where n = 1… length(diag(S)) to identify the number of the principal components, as long as the n's value could make this formula exceed than 0.85, its value could be taken as the number of principal components. The realization codes are as follows:

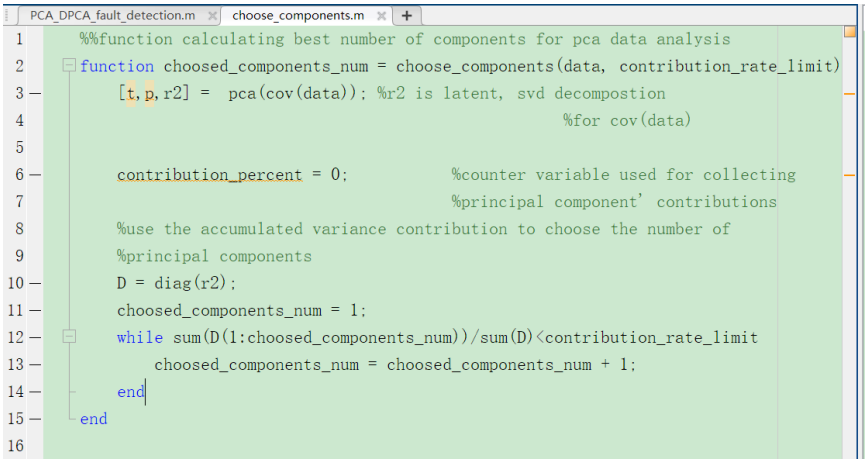


Figure 2.2 codes used for selecting the number of principal components for normal data

Step 3: The next step is to use corresponding formulas to calculate T2 and Q statistics for each sample.

For T2's calculation, we have already have normal data(d00_te) as the training data and fault data(d04_te) as the testing data here, and next is to do the singular value decomposition for (train_data)/sqrt(n-1):

$$\frac{1}{\sqrt{n-1}}\mathbf{X}=\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{2.1}$$

where X is the training data matrix, n is the number of samples in one dataset. U, Sum and V respectively are results of svd decomposition results.

After this procedure, we could use the following formula to calculate the T2 statistic:

$$T^2=\mathbf{x}^T P\boldsymbol{\Sigma}_a^{-2}P^T\mathbf{x} \tag{2.2}$$

where x represents each sample in the dataset, P includes the loading vectors associated with the alpha largest singular values, $\boldsymbol{\Sigma}_a$ contains the first a(the principal components we will take) rows and columns of $\boldsymbol{\Sigma}$.Given a number of loading vectors, alpha(1-alpha is the confidence of T2 statistic), a, the threshold can be calculated for the T2 statistic using the probability distribution.

$$T_\alpha^2=\frac{(n^2-1)a}{n(n-a)}F_\alpha(a,n-a) \tag{2.3}$$

where $F_\alpha(a,n-a)$ is the upper 100 * alpha % critical point of the F-distribution with a and n-a degrees of freedom.

And the realization codes for calculating T2 statistic are as follows:

```
%%T2 calculation
function [T2,T2_threshold] = T2_calculation(train_data,test_data,sample_num,...
    remain_components,alpha)
    %use pca method to get principal components
    [t,p,r2] = pca(train_data);
    s = svd(train_data/sqrt(sample_num-1));
    P = p(:,1:remain_components);
    sigma2 = s(1:remain_components).^(-2);
    for i = 1:size(train_data,1)
        x = test_data(i,:)';
        T2(i) = x'*P.*sigma2'*P'*x;
    end
    %calculate the t2 threshold level
    ts = finv(1-alpha, remain_components, sample_num-remain_components);
    T2_threshold = ((sample_num.^2-1)*remain_components)/(sample_num*...
        (sample_num-remain_components)).*ts;
end
```

For Q statistic's calculation, we used formula developed by Jackson and Mudholkar.

$$Q = r^T r, \quad r = \left(I - PP^T\right)\mathbf{x} \tag{2.4}$$

The threshold for Q statistic can be calculated from its approximate distribution:

$$Q_\alpha = \theta_1\left[\frac{h_0 c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_0\left(h_0 - 1\right)}{\theta_1^2}\right]^{1/h_0}$$

Where $\theta_i = \sum_{j=a+1}^{n} \sigma_j^{2i}, h_0 = 1 - \left(2\theta_1\theta_3\right)/3\theta_2^2$,and c_α is the normal deviate corresponding to the (1-alpha) percentile.

The realization programs for calculating Q statistic and its threshold value are as follows:

```
function [Q,Q_threshold] = Q_calculation(train_data,test_data,sample_num,...
    remained_components,alpha)
    [t0,p0,r2] = pca(train_data);
    s = svd(train_data/sqrt(sample_num-1));

    projected_t = test_data*p0;  %variable data(dataset) is projected to axis of
                                 %d00_te
    %Q2
    Q = test_data - projected_t(:,1:remained_components)*p0(:,1:remained_components)';
    Q = Q';
    Q = sum(Q.^2);

    s = s(remained_components+1:end);
    theta1 = sum(s.^2);
    theta2 = sum(s.^4);
    theta3 = sum(s.^6);

    h0 = 1 - (2*theta1*theta3)/(3*theta2^2);
    ca = norminv(1-alpha);

    %Q threshold
    Q_threshold = theta1*((h0*ca*sqrt(2*theta2))/theta1 + 1 + (theta2*h0*...
        (h0-1))/theta1^2)^(1/h0);
end
```

Step 4: draw figures of T2 and Q statistics for the fault dataset and analyze results

Step 5: Take two abnormal samples from T2 and Q figures, using them to calculate each variables' contribution to these two statistics for judging which variables could be the main causes during the fault happens and using histograms to describe it. Use formulas below to calculate each variables' contributions to T2 and Q statistics:

For each variables' contribution for Q statistic, it could be calculated by using the following formula:

$$\mathrm{Cont}_i^Q = \left(\xi_i^T C \cdot x\right)^2, i = 1, \ldots, m \tag{2.5}$$

where x is the new collected fault data sample which is a m by 1 vector, Cont_i^Q indicates each variable's contribution value for Q statistic, C = I – P*P', ξ_i indicates the i th column of identity matrix I.

For each variables' contribution for T2 statistic, it could be calculated by using the following formula**:**

$$\mathrm{Cont}_i^{T^2} = x^T \cdot D \cdot \xi^i \xi_i^T \cdot x \tag{2.6}$$

where $D = P^T \Lambda^{-1} P$, P is the loading matrix, and Λ is just the Σ in formula 2.1, x, $\xi^i \xi_i^T$ have the same meanings with parameters in formula 2.5.

Step 6: Respectively build the DPCA extended matrices for normal and fault data, and the matrix's build-up way will be like this:

$$\mathbf{X}(l) = \begin{bmatrix} x_t^T & x_{t-1}^T & \cdots & x_{t-l}^T \\ x_{t-1}^T & x_{t-2}^T & \cdots & x_{t-l-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ x_{t+l-n}^T & x_{t+l-n-1}^T & \cdots & x_{t-n}^T \end{bmatrix} \tag{2.7}$$

where $\mathbf{x}_t^T$ is the m-dimensional observation vector is the training set the time instance t. and all columns' data behind the first column represents all the historical data before the time instance t, and the main matlab realization is as follows:

```
%%DPCA method
history_num = 4;    %use the fronter four moments data to spin at the row end
                    %of the matrix
[row,column] = size(normal_data);   %get the row and column info of the matrix
normal_data_dpca_matrix = zeros(row-history_num, (history_num+1)*column);
fault_data_dpca_matrix = zeros(row-history_num, (history_num+1)*column);

%use the before four moments' data as history data,and put the data of
%current moment at the start of each row, so the number of total columns
%DPCA matrix will be (history_num+1)*column),and the number of total rows
%will be row-history_num
for i = history_num+1:row   %i represents the row
    for j = 0:history_num
        normal_data_dpca_matrix(i-history_num,j*column+1:(j+1)*column) = normal_data(i-j,:);
        fault_data_dpca_matrix(i-history_num,j*column+1:(j+1)*column) = fault_data(i-j,:);
    end
end
```

Step 6: Repeat step 1,3,4 operations for extended matrices of normal and fault data.

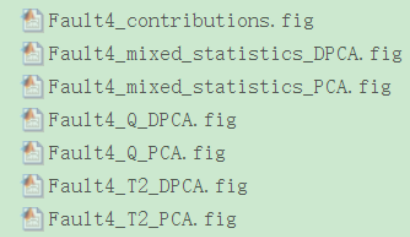
Step 7: For more precisely analyze, use some mixed statistic combining T2 and Q statistics, the formula is as follows:

Mixed statistic = Q values/Q threshold value + T2 values/T2 threshold value

and use this formula to calculate mixed T2 and Q statistics for PCA and DPCA analyzed results.

3. Results and analysis

Totally we have used three fault datasets: d04_te.dat, d11_te.dat, d15_te.dat, and for each datasets, there are seven figures generated from program running results, they respectively are:T2 and Q statistics' figures using PCA method, T2 and Q statistics' figures using DPCA method, Mixed statistic figures for PCA and DPCA methods , and each variables contribution histogram for T2 and Q statistics, let's take fault dataset d04 as the example, the figures' file structure is as follows:



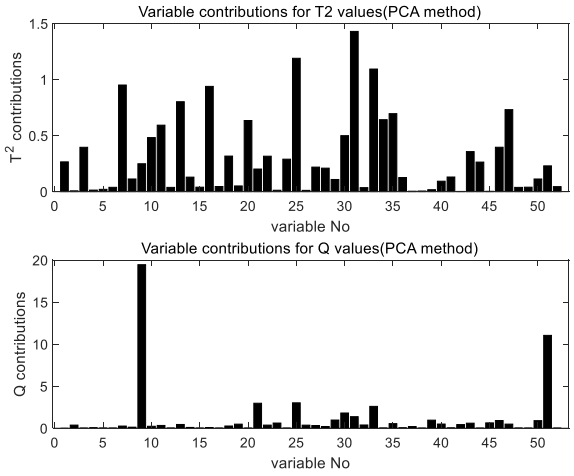
And for drawing histograms when some fault happens, we build up one matrix recording one random abnormal sample' fault positions for T2 and Q statistics at each dataset, and the matrix is as follows:

```
pca_error_positions = [161, 165; 356, 167; 100, 293];
```

Each row's data represents the fault happened location observed from T2 and Q figures for each dataset.

Next we will respectively analyze figures generated by matlab programs for three datasets:

For d04_te, the contribution histogram is like this (For T2 histogram, we take fault position 161 to calculate the T2 contribution, For Q histogram, we take fault position

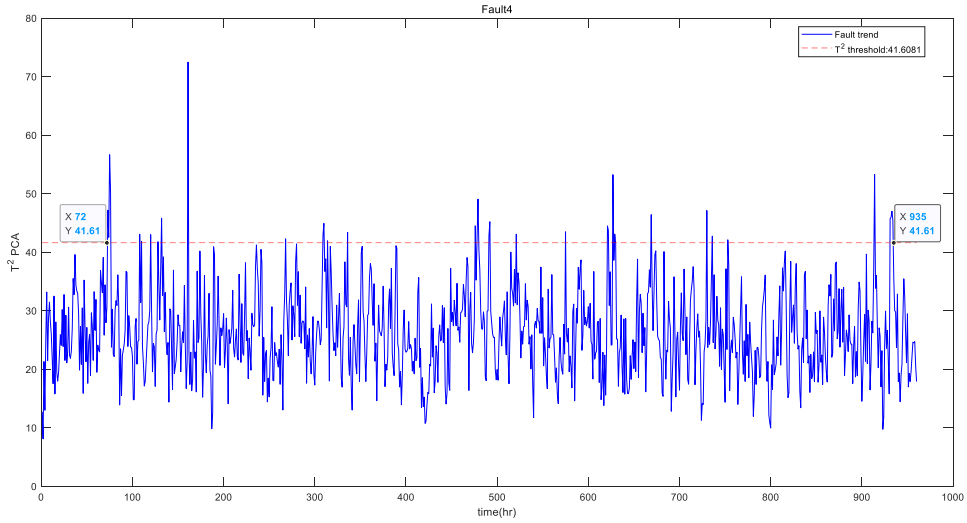


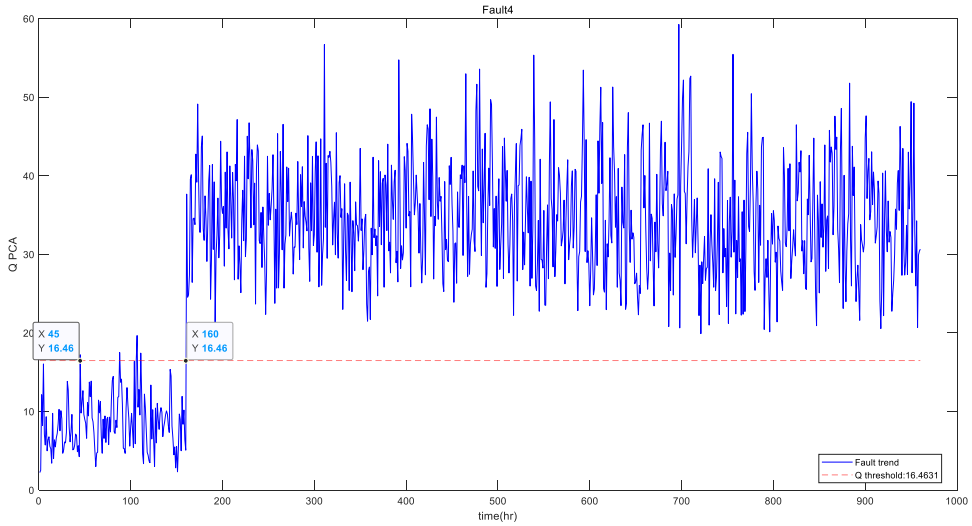
165) :

From figures above, we could know the variable 31 make the greatest contribution for T2 statistic when the fault happens at the position 161(time instance), which means that variable 31 could be the main cause of the time 161 fault at the T2 figure, but for Q statistic, variable 9 make the greatest contribution when the fault happens at the position 165,which also indicates that variable 9 could be the main cause of the time 165 fault at the Q figure.

Note: Because sometimes fault happened pretty frequent at T2 and Q's statistic figure, so for displaying conveniently reasons, we only pick random abnormal samples spotted at T2 and Q statistic figures to analyze which variables has caused the fault.

The T2 and Q statistic figure of fault data for d04_te are like this:





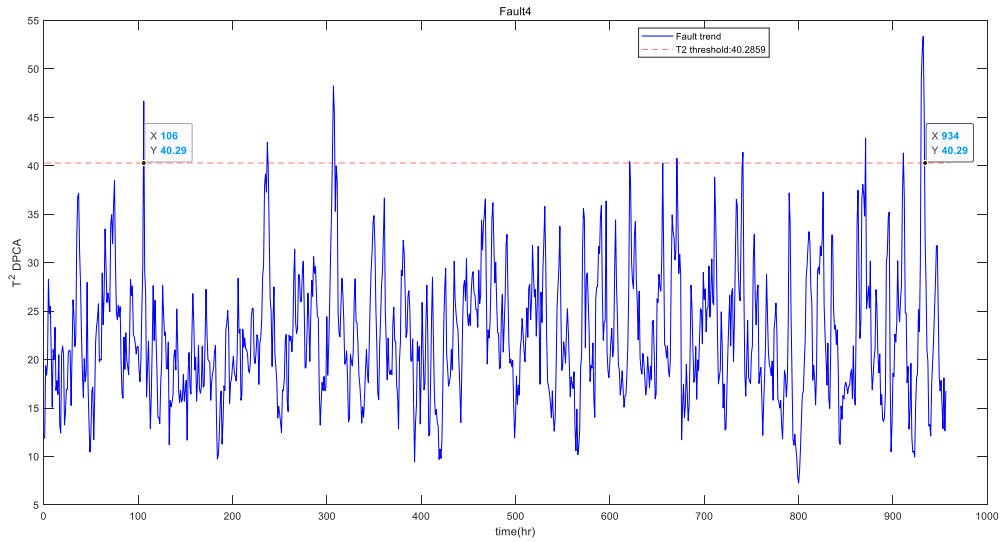
For T2 and Q calculated by using DPCA method

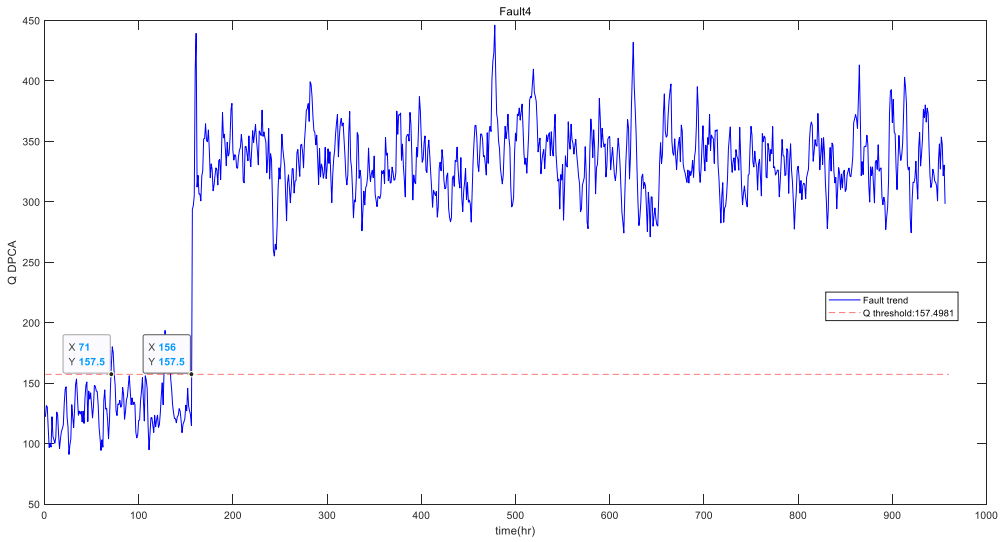
the T2 statistic indicates that T2 enter the abnormal status at moment 72, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 935, the T2 statistic totally gets back to the normal again.

For Q statistic figure of fault4, when the curve gets to the position 45, the fault happened, and when the curve gets to the moment 160, one obvious jump happened at

the curve, and after that the curve just stays at the abnormal status and never get back to the normal again.

For T2 and Q calculated by using DPCA method, figures are as follows:

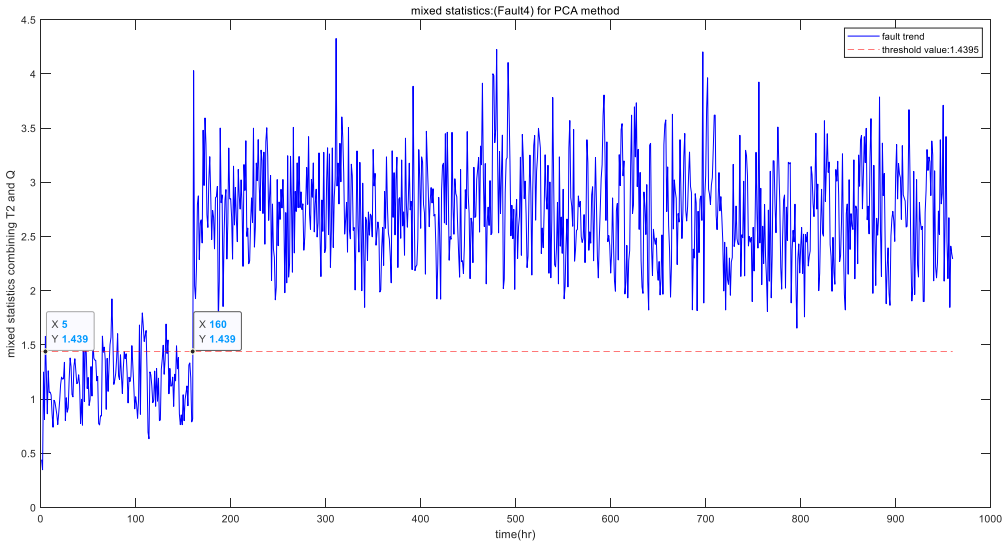


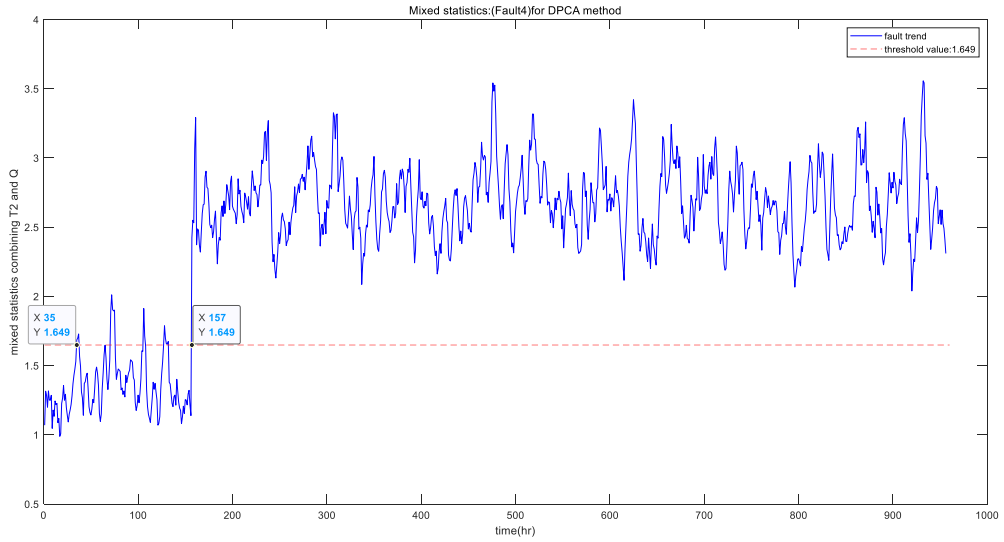


The T2 statistic indicates that T2 enter the abnormal status at moment 106, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 934, the T2 statistic totally gets back to the normal again.

For Q statistic figure of fault4, when the curve gets to the position 71, the fault happened, and when the curve gets to the moment 156, one obvious jump happened at the curve, and after that the curve just stays at the abnormal status and never get back to the normal again.

For mixed statistic combining T2 and Q calculated by using PCA and DPCA method, figures are as follows:



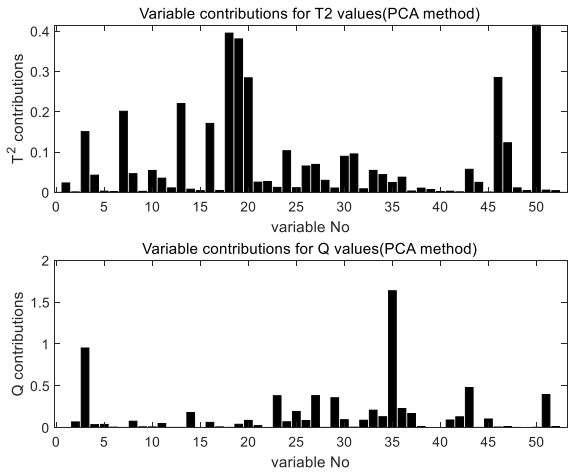


For Mixed statistic calculated by using PCA method's results (T2 and Q), the abnormal starts at moment 5, and at moment 160 there is an obvious jump of the curve, the whole system totally enter the abnormal status and never get back to the normal after then.

For Mixed statistic calculated by using DPCA method's results (T2 and Q), the abnormal starts at moment 35, and at moment 157 there is an obvious jump of the curve, the whole system totally enter the abnormal status and never get back to the normal after then.

For d11_te, the contribution histogram is like this (For T2 histogram, we take fault position 356 to calculate the T2 contribution, For Q histogram, we take fault position

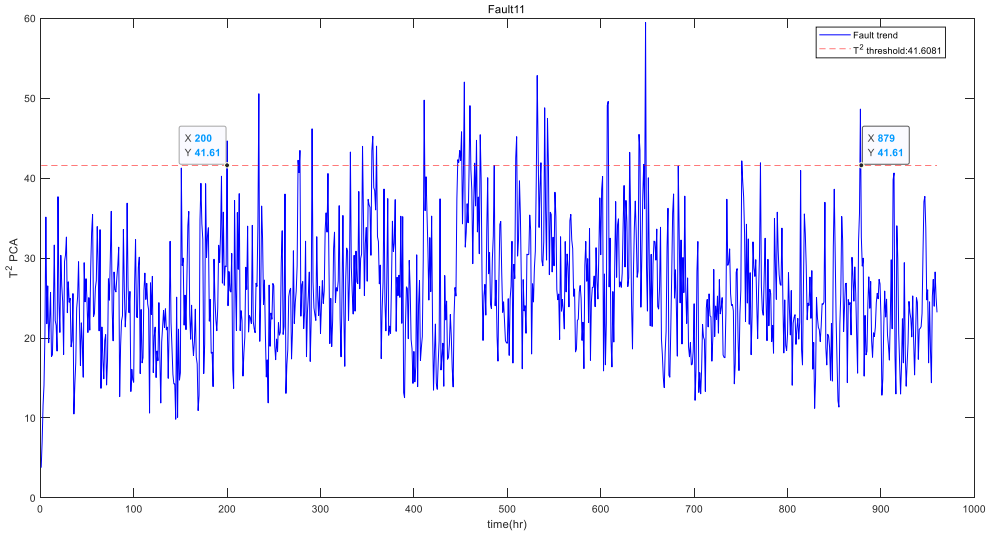
167) :

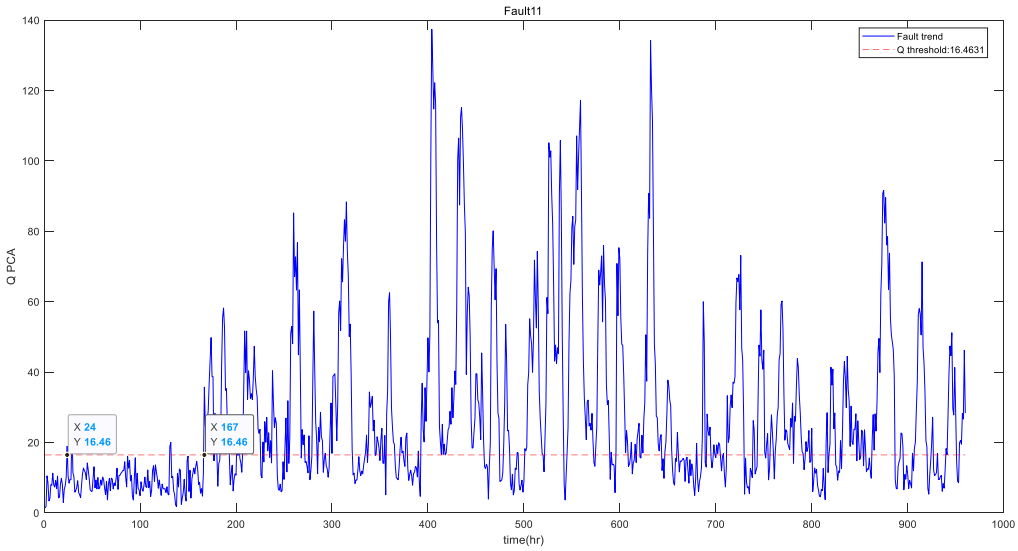


From figures above, we could know the variable 50 make the greatest contribution for T2 statistic when the fault happens at the position 356(time instance), which means that variable 50 could be the main cause of the time 356 fault at the T2 figure, but for Q statistic, variable 9 make the greatest contribution when the fault happens at the position 167,which also indicates that variable 35 could be the main cause of the time 167 fault at the Q figure.

The T2 and Q statistic figure of fault data for d11_te are like this:

For T2 and Q calculated by using PCA method, figures are as follows:

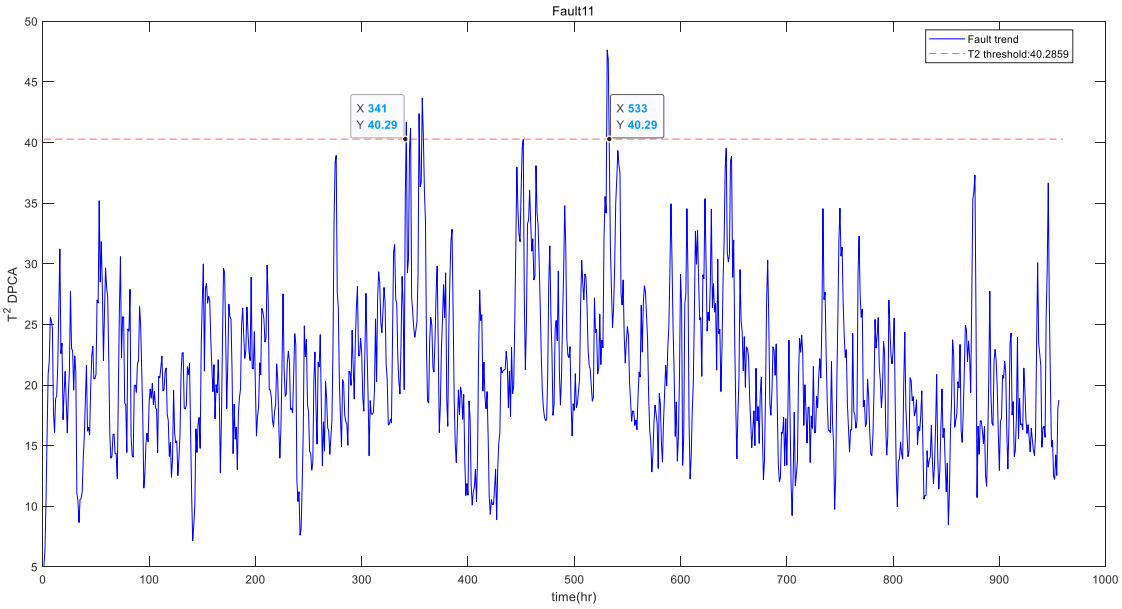


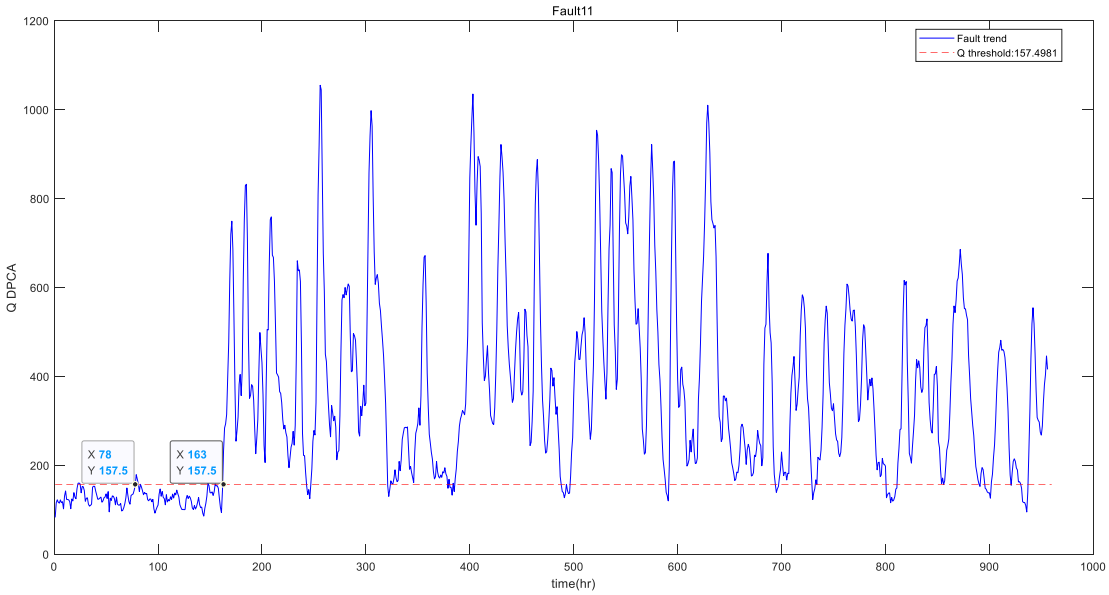


The T2 statistic indicates that T2 enter the abnormal status at moment 200, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 879, the T2 statistic totally gets back to the normal again.

For Q statistic figure of fault11, when the curve gets to the position 24, the fault happened, and when the curve gets to the moment 167, the whole system gets to a extremely unstable status, Q data get back and forth from the normal to the abnormal, and never get to the status staying always at the normal status.

For T2 and Q calculated by using DPCA method, figures are as follows:

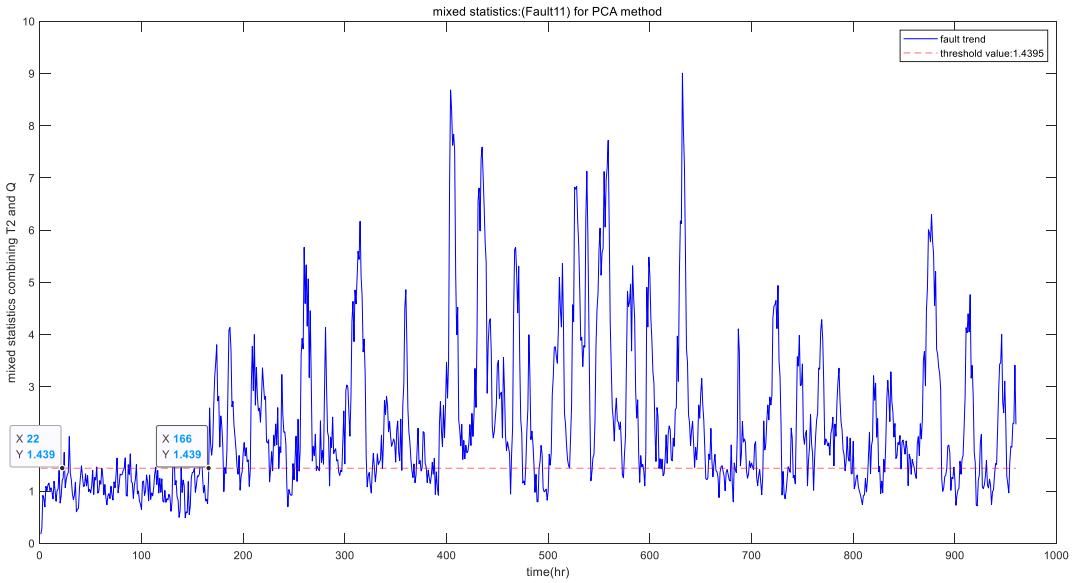


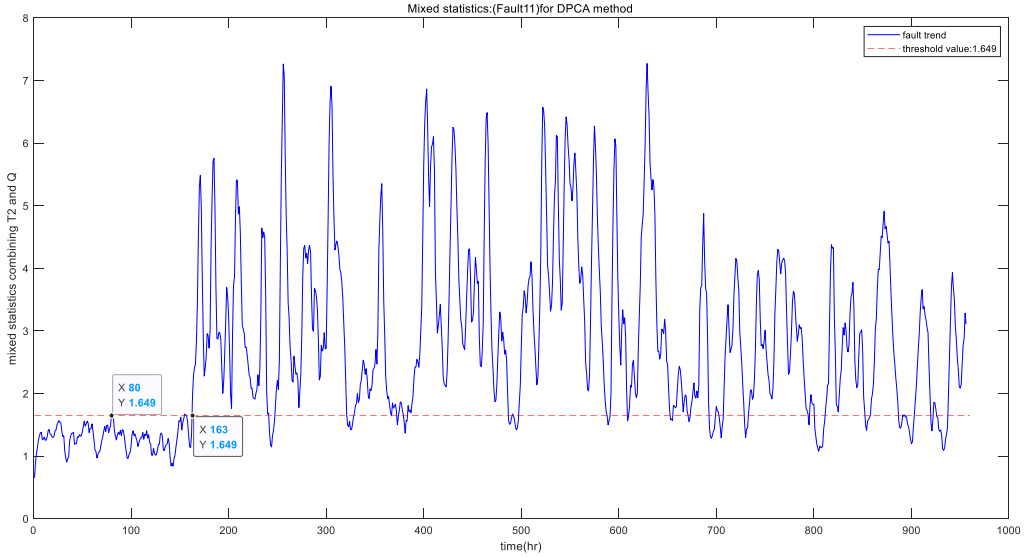


The T2 statistic indicates that T2 enter the abnormal status at moment 341, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 533, the T2 statistic totally gets back to the normal again.

For Q statistic figure of fault11, when the curve gets to the position 78, the fault happened, and when the curve gets to the moment 163 one obvious jump happened at the curve, and after that the curve almost just stays at the abnormal status and never get back to the normal again.

For mixed statistic combining T2 and Q calculated by using PCA and DPCA method, figures are as follows:

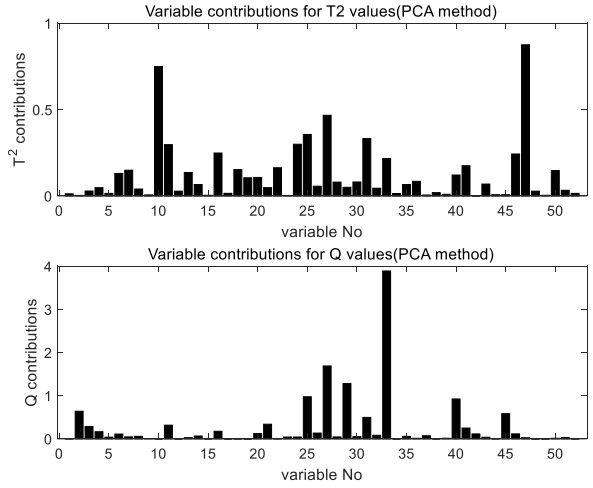




For Mixed statistic calculated by using PCA method's results (T2 and Q), the abnormal starts at moment 22, and at moment 166 there is an obvious jump of the curve, the whole system almost totally enter the abnormal status and sometimes will get back to the normal after then, there are multiple obvious vibrations of this curve during this period.

For Mixed statistic calculated by using DPCA method's results (T2 and Q), the abnormal starts at moment 80 but not very big, and at moment 163 there is an obvious jump of the curve, the whole system almost totally enter the abnormal status and sometimes get back to the normal after then.

For d15_te, the contribution histogram is like this (For T2 histogram, we take fault position 100 to calculate the T2 contribution, For Q histogram, we take fault position 293) :

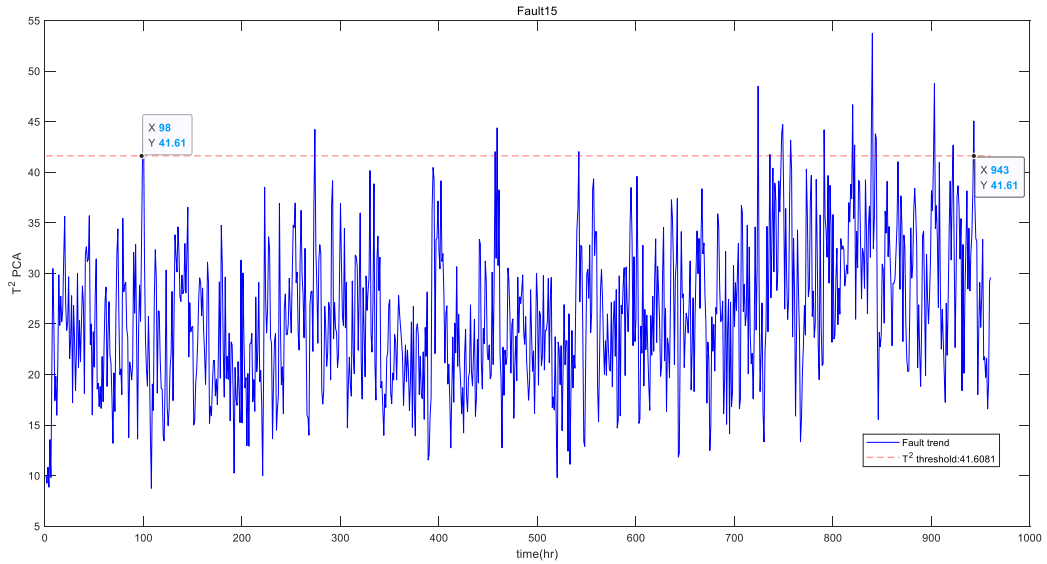


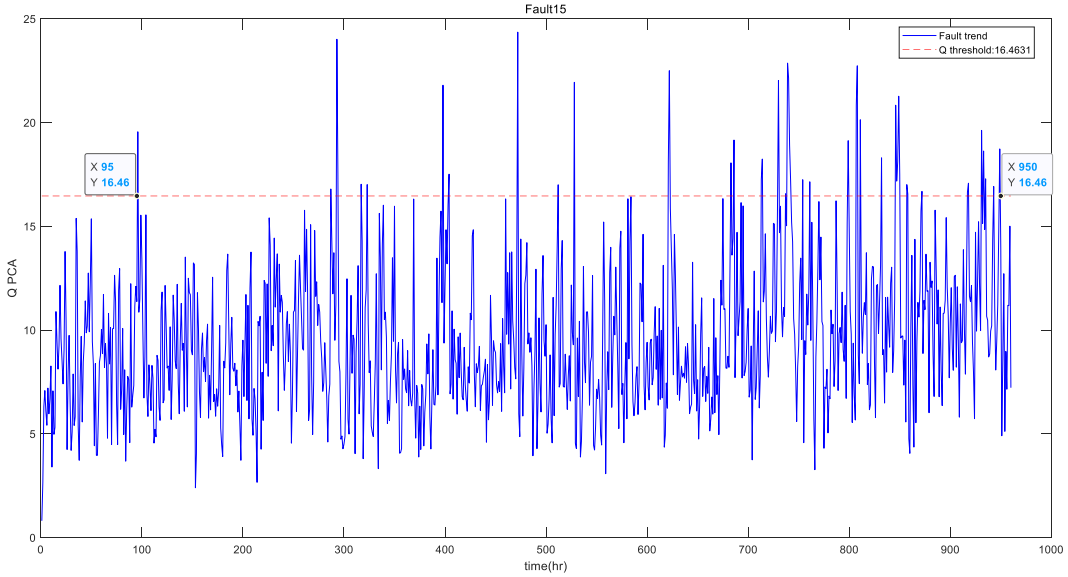
From figures above, we could know the variable 47 make the greatest contribution for T2 statistic when the fault happens at the position 100(time instance), which means that variable 47 could be the main cause of the time 100 fault at the T2 figure, but for Q statistic, variable 33 make the greatest contribution when the fault happens at the

position 293,which also indicates that variable 33 could be the main cause of the time 293 fault at the Q figure.

The T2 and Q statistic figure of fault data for d15_te are like this:

For T2 and Q calculated by using PCA method, figures are as follows:

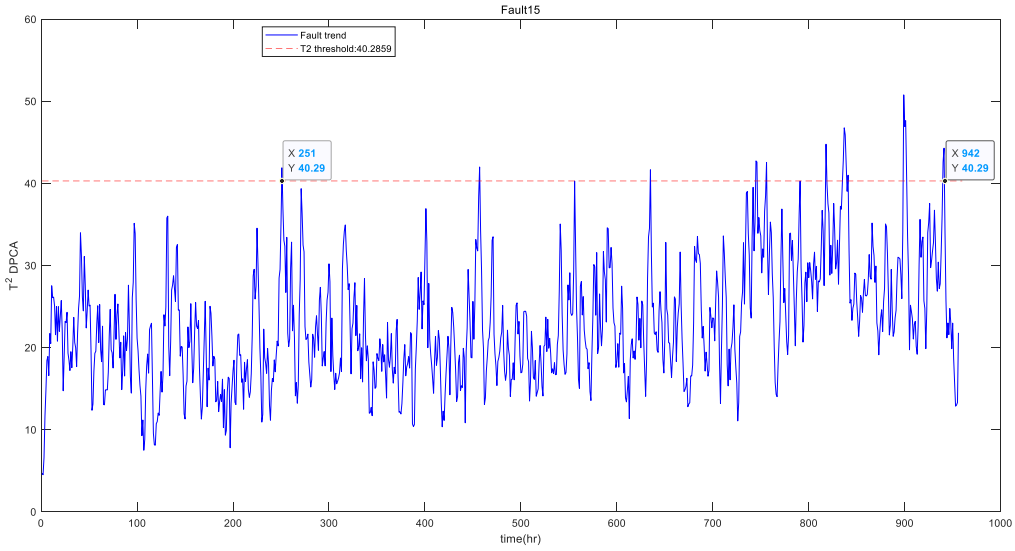


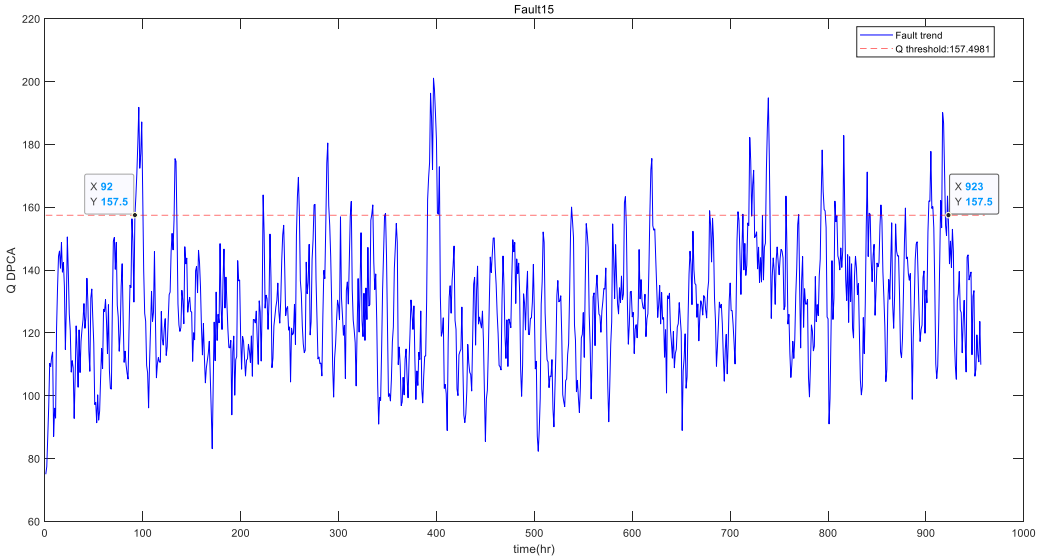


The T2 statistic indicates that T2 enter the abnormal status at moment 98, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 943, the T2 statistic totally gets back to the normal again.

For Q statistic figure of fault15, which enters the abnormal status at moment 95, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 950, the Q statistic totally gets back to the normal again.

For T2 and Q calculated by using DPCA method, figures are as follows:

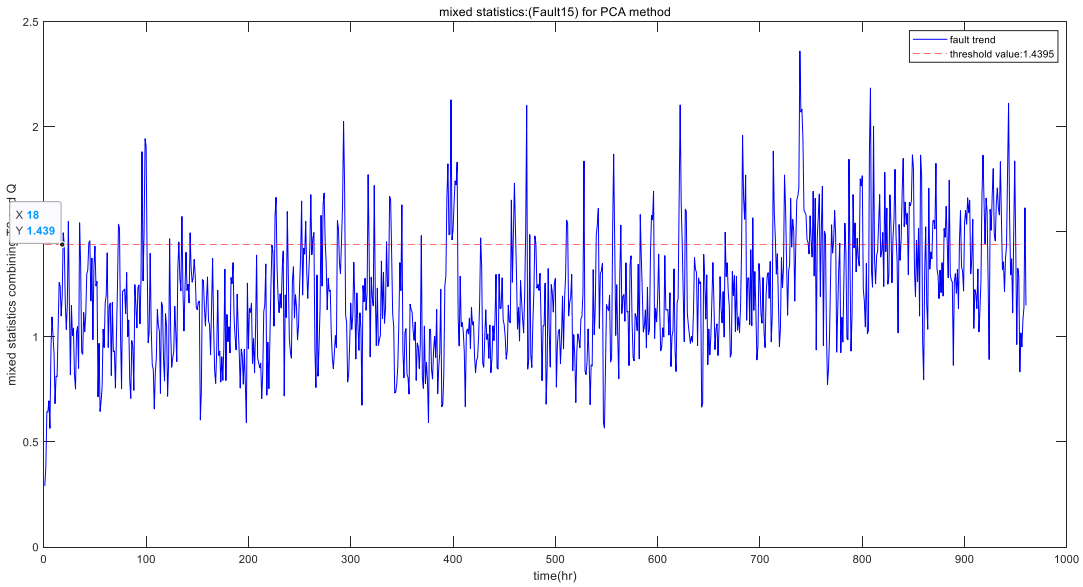


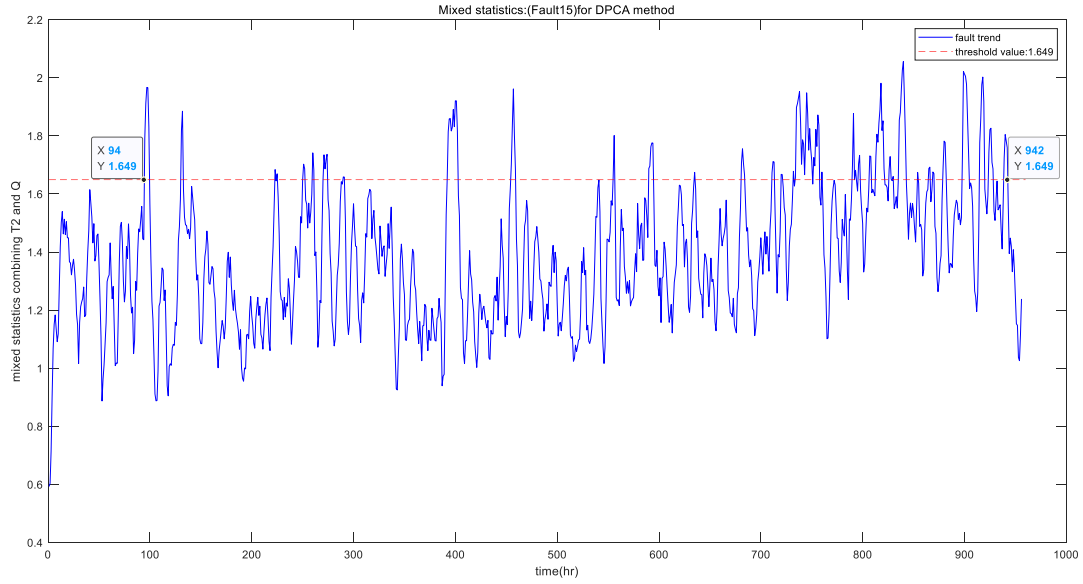


The T2 statistic indicates that T2 enter the abnormal status at moment 251, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 942, the T2 statistic totally gets back to the normal again.

For Q statistic figure of fault15, which enters the abnormal status at moment 92, after that, there are some small sharps (abnormal status) happened, but when the curve gets to the moment 923, the Q statistic totally gets back to the normal again.

For mixed statistic combining T2 and Q calculated by using PCA and DPCA method, figures are as follows:





For Mixed statistic calculated by using PCA method's results (T2 and Q), the abnormal starts at moment 18, and after that the curve has always been staying back and forth between the normal and abnormal level, never has always been staying at the normal level again.

For Mixed statistic calculated by using DPCA method's results (T2 and Q), the abnormal starts at moment 94, and after that the curve has always been staying back and forth between the normal and abnormal level, and getting back to the normal at moment 942.