The Normalization PCA Model and Its Application in Fault Detection of Wind Power Generation System

Wang Tianzhen
Department of Electrical Automation,
Shanghai Maritime University
Shanghai, China
wtz0@sina.com

Tang Tianhao
Department of Electrical Automation,
Shanghai Maritime University
Shanghai, China
thtang@shmtu.edu.cn

Xu Man
Department of Electrical Automation,
Shanghai Maritime University
Shanghai, China
xumanaa_2006@126.com

Han Jingan
Department of Electrical Automation,
Shanghai Maritime University
Shanghai, China
jghan.cn@gmail.com

***Abstract*—The principal component analysis method is usually used for fault detection under the steady conditions, however, when system works under the non-steady conditions, the false alarm rate and the missing alarm rate, tested by the T^2 control limit, are so high. The main reason for this situation is that the sampled data is accord with normal distribution under the steady conditions, whereas the data does not satisfy normal distribution under the non-steady conditions, but T^2 control limit can only detect fault effectively for the data conforming to normal distribution. So this paper firstly analyzes the characteristics of the periodic non-steady conditions and then puts forward a normalization PCA (NPCA) model according to the precondition of effective detection under the T^2 control limit. This model deals with the measured data for normalization data based on the longitudinal standardization, and then uses T^2 control limit to detect fault. At the end, it is applied to wind power generation system, and the results verify the effectiveness of the model.**

Keywords—PCA; T^2 control limit; NPCA; longitudinal standardization; fault detect

I. INTRODUCTION

Over the past 20 years, the rapid of development of computer system and database system has lead to the growth of application requirement with using the multivariate statistical analysis method to deal with data information, and the purpose is to obtain useful information through data analysis to reveal and reflect the change of the internal process [1].

The two most commonly used statistical methods are principal component analysis (PCA) and partial least squares (PLS) [2,3], the two exploited to reduce historical process database in order to have better understanding of hidden database in order to have better understanding of hidden information concealed in data, but they are generally suitable for steady working conditions. Recursive PCA [4] and exponentially weighted PCA (EWPCA) [5] can monitor time-varying industrial process. If the complex system is under the non-steady conditions (The non-steady conditions are referred to such motion process as starting, braking and other mutation conditions [6]), the above method fail to detect fault effectively and computational load is large. The main reason for this phenomenon is that the sample data is accord with normal distribution under the steady conditions, whereas the data does not satisfy normal distribution under the non-steady conditions. However, the actual industrial process is often unsteady, such as wind power generation system [7], because of the season, air pressure and topography influence, the speed of wind changes random and unsteady so significantly that the role of the wind vane is also random and unsteady. In addition, the failure probability of the system under the non-steady conditions is relatively high. For the above problem, a normalization PCA model is proposed. This model can make the data of the non-steady conditions conform to normal distribution [8].

In this paper, PCA method is introduced simply in Section 2 and the T^2 control limit is also given. Section 3 describes the characteristics of the periodic non-steady conditions in detail. This section firstly analyze the characteristics of the periodic, then some properties are proposed and the concept of longitudinal standardization is defined, finally mathematical proof and Q-Q plots commonly prove these properties. Section 3 has laid the theoretical foundation for the new model. The proposed model is introduced in Section 4 and this section mainly gives the process of data normalization and the special steps of fault detection based on NPCA. This model is applied to the wind power generation system under the non-steady conditions to test the validity of the method and discuss its test results in Section 5.

II. PCA -BASED FAULT DETECTION

PCA is a multivariate statistical method for classification, which has been widely used in fault diagnosis fields in recent years. A reduced representation of original data is obtained which is smaller in size but having enough information to deal

Supported by the project of NSFC (NO.61203089)

with. These variables are ordered regarding their importance and the magnitude of each eigenvalues describes the variance in the direction of its corresponding eigenvector. So the importance of each variable can be considered with its eigenvalues [9,10].

A. PCA Method

Suppose X is an $n \times N$ data matrix of system variable sequence.

$$X = [x(1), x(2), \cdots, x(N)] \tag{1}$$

$$x(k) = [x_1(k), x_2(k), \cdots, x_n(k)]^T, k = 1, 2, \cdots, N \tag{2}$$

Where n is the number of variables, N is the number of samples, and X is decomposed as follows:

$$X = \sum_{i=1}^{n} p_i t_i = \sum_{i=1}^{m} p_i t_i + E \tag{3}$$

Where $t_i \in R^{1\times N}$ are the score matrix and $p_i \in R^{n\times 1}$ are the loading matrix, $m(m < n)$ is the number of principal components. Respectively, the PCA projection reduces the original set of n variables to m principal components.

Cumulative percent variance (CPV) is a measure of the percent variance captured by the m PCs:

$$CPV(m) = \left(\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \right) \times 100\% \geq CL \tag{4}$$

The number of PCs is chosen when the value of CL is determined by the user as predetermined limit.

B. T^2 Control Limit Based on PCA

For fault detection, T^2 statistics is often used. The T^2 statistics is a measure of the variation in principal component space:

$$T^2(k) = \sum_{i=1}^{m} \frac{t_i(k)^2}{s_{t_i}^2} \tag{5}$$

Where $t_i(k)$ is the value of the i th row of the score matrix t_i, $s_{t_i}^2$ is the estimation variance of t_i.

The T^2 control limit can be calculated by using F distribution, and its equation is as follows:

$$UCL = \frac{m(N-1)}{N-m} F_\alpha(m, N-m) \tag{6}$$

Where, UCL is the upper limit of T^2 statistics, N is the number of samples, and m is the number of reserved PCs. Confidence $1-\alpha$ could be determined by the user need, for example, if confidence is 95% or 99%, then $\alpha = 0.05$ or 0.01. $F_\alpha(m, N-m)$ is the critical value of F distribution corresponding to the test level α, the degree of freedom m and $N-m$. If the process is abnormal, then $T^2 > UCL$.

III. Characteristics Of The Periodic Non-Steady Conditions

The new model proposed in this paper is suitable for periodic non-steady conditions, so characteristics of the periodic non-steady conditions must be provided relevant explanation before the model is constructed. This section firstly analyzes some characteristics about the periodic non-steady conditions and proposes three properties, then gives the simple mathematical proof, finally uses Q-Q plots to assess the assumption of normality for further verification [11].

A. Characteristics Analysis of the Periodic Non-steady Conditions

The non-steady conditions can be classified the periodic non-steady conditions and the non-periodic non-steady conditions. The periodic non-steady conditions refer to such non-steady conditions that system runs in accordance with a series of same operating circles and usually is divided into two types: One kind of period is in the strict sense, namely the internal environment and the external working conditions of the system in different periods are the same; Another kind is to point to the broad sense, namely some main conditions are continuously repeated, consistent with periodic, but some minor conditions do not satisfy the periodic can also be approximately considered to conform to the periodic. For example, a plastic bag machine servo motor works under the periodic non-steady conditions, and its motion includes four processes: acceleration, constant speed, deceleration, speed recovery. The periodic is in the strict sense in the paper.

Property 1: The measured data of different time at the same period do not conform to the normal distribution under the periodic non-steady conditions.

Property 2: The measured data of different periods at the same time do conform to the normal distribution under the periodic non-steady conditions.

Proof:

Suppose $X = [x(1) \quad x(2) \quad \cdots \quad x(N)]$ is a set of periodic data, and $x(k + (\beta - 1)N) = x(k)$, $\beta = 1, 2, \cdots$, $k = 1, 2, \cdots, N$, N is the period.

In actual industrial process, all the measured data exist error ς, and the errors are accord with normal distribution $N(0, \chi)$. Suppose the theoretical true value is ψ, so the data can be described as $x(k + \beta N) = \psi(k) + \varsigma(k + \beta N)$. Take j periods and we can get a set of data

$$\begin{aligned} X(k) &= [x(k), x(k+N), \cdots, x(k+jN)] \\ &= [\psi(k) + \varsigma(k), \psi(k) + \varsigma(k+N), \cdots, \\ &\quad \psi(k) + \varsigma(k+jN)] \end{aligned} \tag{7}$$

Then $\begin{matrix} X(k)-[\psi(k)\ \psi(k)\ \cdots\ \psi(k)] \\ =[\varsigma(k)\ \varsigma(k+N)\ \cdots\ \varsigma(k+jN)] \end{matrix}$ conforms to the normal distribution $N(0,\chi(k))$, so $X(k)$ conforms to the normal distribution $N(\psi(k),\chi(k))$ and the proof of the property 2 has been given.

The change amplitude of sample data is very large, so $\exists\delta\in[1,2,\cdots,N]$ make $\psi(\delta)\neq\psi(k)$. The proof of the property 1 has been given.

Definition: Longitudinal standardization

When j is big enough, the mean value $\overline{X}(k)$ of $X(k)$ can be approximately equal to $\psi(k)$, namely $\overline{X}(k)\approx\psi(k)$ and the standard deviation $S(k)\approx\chi(k)$. Standardize $X(k)$ to get $Y(k)=[y(k),y(k+N),\cdots,y(k+jN)]$.

$$y(k+\beta N)=\frac{x(k+\beta N)-\overline{X}(k)}{S(k)} \tag{8}$$

This process can be defined as longitudinal standardization.

Property 3: The measured data of different time at the same period do conform to the standard normal distribution after longitudinal standardization under the periodic non-steady conditions.

Because the longitudinal standardization data $Y(k)$ conforms to the standard normal distribution $N(0,1)$, $Y=[y(1)\ \ y(2)\ \ \cdots\ \ y(N)]$ conform to the standard normal distribution $N(0,1)$.

B. Normality Assumption Testing

Plots are always useful devices in any data analysis. Special plots called Q-Q plots can be used to assess the assumption of normality. They are, in effect, plots of the sample quantile versus the quantile one would expect to observe if the observations actually were normally distributed. When the points lie very nearly along a straight line, the normality assumption remains tenable. Normality does not exist if the points deviate from a straight line [12].

For a standard normal distribution, the quantiles are defined by the relation

$$P[Z\leq q(j)]=\int_{-\infty}^{q(j)}\frac{1}{2\pi}e^{-z^2/2}dz=\frac{j-0.5}{N} \tag{9}$$

The steps leading to a Q-Q plot are as follows:

1) Order the original observations to get $z(1),z(2),\cdots,z(N)$ and their corresponding probability values $\frac{1-0.5}{N},\frac{2-0.5}{N},\cdots,\frac{N-0.5}{N}$;

2) Calculate the standard normal quantiles $q(1),q(2),\cdots,q(N)$;

3) Plot the pairs of the observations $(q(1),z(1)),(q(2),z(2)),\cdots,(q(N),z(N))$, and examine the "straightness" of the outcome.

This paper adopts the data of the wind power generation system (The introduction of the model and the selection of data will be described in detail in section 4), and uses Q-Q plots to verify properties of section 3.1. Take one group of variable data, i.e., A phase current as an example. Take 400 groups of historical periodic data, and the cycle length is 800.

In Fig.1, (a) is the Q-Q plot of one period of data of A phase current, from which we can see the distribution of data points is in curve-shaped, does not meet the linearity, so this group of data does not satisfy the normal distribution which could explain property 1. (b) and (c) are Q-Q plots for the moment k=1, 400 about 400 different periods of data, we can see that the distribution approximates a straight line and meet the linearity, so the data of different periods of the two moment conforms to normal distribution. The data of other moments can be verified meeting the normal distribution using the same method, so property 2 can be verified. (d) is the Q-Q plot of one period of data after normalization, and the data distribution approximates a straight line, meet the linearity, so property 3 has been verified.

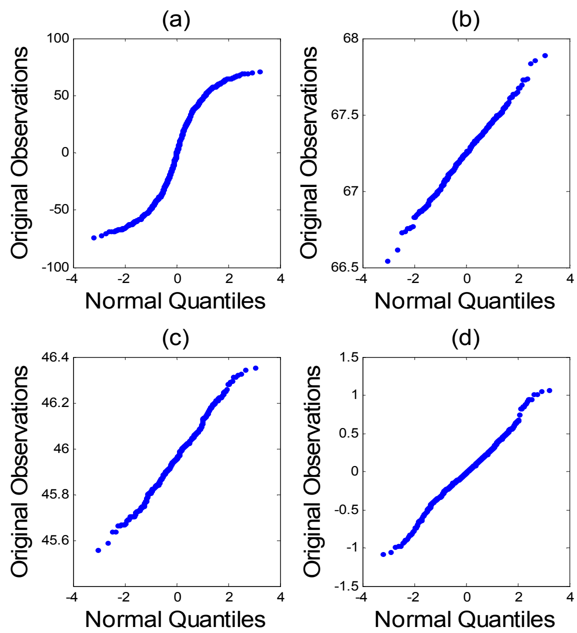

Figure 1. The Q-Q plots of four cases

IV. THE NPCA MODE

Section 3 has analyzes the applicable conditions by using T^2 control limit for fault detection under the periodic non-steady conditions and has proposed some properties, thus we can construct the NPCA model and the model mainly includes three steps: data normalization, PCA and fault detection based on T^2 control limit. This section will introduce the three steps in detail next.

A. Data Normalization

Sample a large number of historical data for mean value and variance, and then standardize the measured data according to longitudinal standardization. The specific steps are as follows:

1) Sampling historic data

In the historical data set, collect j groups of historical data $X_1^H, X_2^H, \cdots X_j^H$, the cycle length is N and the number of variables is n.

Where $$X_\beta^H = \begin{bmatrix} x_{\beta 1}^H(1) & x_{\beta 1}^H(2) & \cdots & x_{\beta 1}^H(N) \\ x_{\beta 2}^H(1) & x_{\beta 2}^H(2) & \cdots & x_{\beta 2}^H(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_{\beta n}^H(1) & x_{\beta n}^H(2) & \cdots & x_{\beta n}^H(N) \end{bmatrix},$$ $\beta = 1,2,\cdots,j$

2) Calculating mean value and variance

According to the j groups of historical data, calculate the mean value and variance of each moment of each variable, i.e.

$$\overline{X}^P = \begin{bmatrix} \bar{x}_1^P(1) & \bar{x}_1^P(2) & \cdots & \bar{x}_1^P(N) \\ \bar{x}_2^P(1) & \bar{x}_2^P(2) & \cdots & \bar{x}_2^P(N) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_n^P(1) & \bar{x}_n^P(2) & \cdots & \bar{x}_n^P(N) \end{bmatrix}$$

and $$S^P = \begin{bmatrix} s_1^P(1) & s_1^P(2) & \cdots & s_1^P(N) \\ s_2^P(1) & s_2^P(2) & \cdots & s_2^P(N) \\ \vdots & \vdots & \ddots & \vdots \\ s_n^P(1) & s_n^P(2) & \cdots & s_n^P(N) \end{bmatrix}.$$

3) Sampling real-time data online

Sample a group of measured data $$X_{test} = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(N) \\ x_2(1) & x_2(2) & \cdots & x_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(1) & x_n(2) & \cdots & x_n(N) \end{bmatrix}$$ online.

4) Longitudinal standardization

Standardize data X_{test} longitudinally to get

$$X_{test}^P = \begin{bmatrix} x_1^P(1) & x_1^P(2) & \cdots & x_1^P(N) \\ x_2^P(1) & x_2^P(2) & \cdots & x_2^P(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_n^P(1) & x_n^P(2) & \cdots & x_n^P(N) \end{bmatrix},$$

where $x_i^P(k) = \dfrac{x_i(k) - \bar{x}_i^P}{s_i^P}, i = 1,2,\cdots,n; k = 1,2,\cdots,N$.

The above four processes are called data normalization and data normalization mainly make the measured data conform to normal distribution, thus T^2 control limit can detect fault effectively.

B. Extracting PCs based on PCA

PCA method is for data dimension reduction but can reserve a lot of important information, mainly include extracting PCs and calculating T^2 statistics. The main steps are as follows:

1） Calculate covariance matrix $R_{X_{test}^P}$ of X_{test}^P.

2） Calculate the eigenvalue λ and the corresponding eigenvector p of $R_{X_{test}^P}$, and the number m of PCs is decided by user need.

3） Calculate T^2 statistics according to Eqn.(5).

C. Fault Detection Based on T^2 Control Limit

T^2 control limit can detect fault effectively after the above two steps about data processing and calculation, and the detection process is:

1) Calculate T^2 control limit UCL according Eqn.(6).
2) Detect whether T^2 statistics of X_{test}^P is over the T^2 control limit UCL, if it is true, then output the fault time and make the corresponding action according to the system requirements.
3) Otherwise, return to STEP1 to sample data online and go on fault detection of the next process.

So the flow chart of fault detection based on this model is Fig.2.

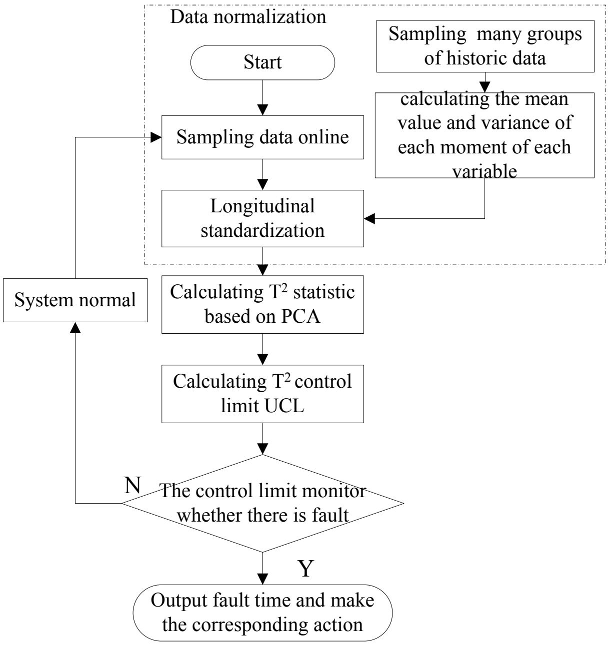

Figure 2. Flow chart of fault detection based on NPCA

V. APPLICATION OF THE NPCA IN WIND POWER GENERATION

Wind power is a pollution-free renewable energy, which is exhaustless and wide distribution. With the requirement of ecological environment and energy, the development of wind energy is taken seriously increasingly. The wind power generation will be large-scale development in the 21st century [13].

Direct-drive permanent-magnet synchronous generation (DDPM) is a main direction of wind power generation [14]. But the permanent magnet material requires high stability, and the weight of the motor increases, as same as that the capacity of the inverter becomes larger. Those shortages lead that the cost of the generator is high. So generator fault could cause great economic loss, so it is very important to detect fault of DDPM. Because of those reasons, we choose DDPM as fault detection model in this paper. The Fig.3 shows the structure of DDPM. The system is composed of wind turbine, rectifier, inverter and maximum power point tracking (MPPT).

Wind speed changes the randomness and instability significantly, so the action in the wind vane is random and uncontrollable, and there are a large amount of unstable factors. According to the mathematical formula $P=\frac{1}{2}\rho\pi R^2 C_p(\lambda)V^3$ (ρ is the air density, R is radius of the rotor, λ is the tip-speed ratio, $C_p(\lambda)$ is the utilization coefficient of wind power, which is related with tip velocity ratio. V is the wind speed) ,it is to build the model of direct-drive permanent-magnet synchronous generation. In order to maximize absorption wind, wind turbines always run in the maximum power point and generator system output power must match wind turbines capture mechanical power strictly. In Normal work, wind speed is among $6\,m/s$ and $11\,m/s$, and different wind turbines cut and cut out at different wind speed. Table1 is the main parameters of the wind generator power. Assume that the change of wind speed is cyclical and the length of every time sampling is as same as the one of wind speed, then get a series of cycle sampling data.

Suppose wind speed is normal at the range $6m/s-11m/s$, run the simulation model many times, get 400 groups of normal fan parameters data (speed, voltage, power, three-phase rotor current, etc.), and calculate the mean and variance which are required in the method of longitudinal standardization. The cycle size is 800, and the confidence of T^2 control limit is 95%.

In sampling time 200-400, the wind speed of the simulation model take about $14m/s$ randomly. Wind speed of the fan in other time is random, but in normal wind speed range. When the wind speed exceeds $11m/s$, the fan withstand mechanical stress is greater than the rated maximum stress, and the long time operation will damage the wind turbine. So the data which is got under above condition is as a group of fault data.

The false alarm rate and the missing alarm rate are important parameters to measure whether a fault detection model is reliable or not, and is also the main basis of feasibility verification. Fig.4 and Fig.5 are the fault detection result of wind power generation system based on PCA and the NPCA. Table 2 lists the false alarm rate and the missing alarm rate of two kinds. Comparing the two methods, the result is that the method based on PCA could not effectively detect the fault, exists seriously missing alarm phenomenon, and detection sensitivity is low. Moreover, the NPCA could effectively overcome the defect of the PCA method under the non-steady conditions. It makes the system keep the lower missing alarm rate and false alarm rate, and enhances the effectiveness of the system monitor greatly.

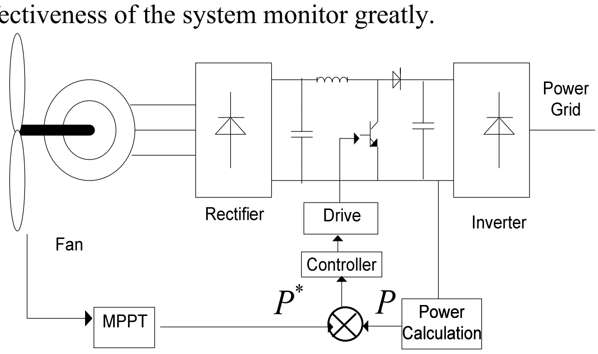

Figure 3. The model of direct-drive permanent-magnet synchronous generator

TABLE I. THE MAIN PARAMETERS OF THE WIND GENERATOR POWER FEEDBACK MODEL

Mechanical (W)	39900	Rotor flux (Wb)	0.192
Generator power (V/A)	39900/0.9	Friction coefficient ($N.m.s$)	0.001889

Pitch angle($^{\circ}$)	0	The optimal tip speed ratio	8.1
Stator resistance (Ω)	0.05	Fan radius R (m)	15
Inductance (H)	0.000635	Wind speed range (m/s)	6-11
Pole logarithmic (P)	36	The biggest wind power utilization coefficient	0.48

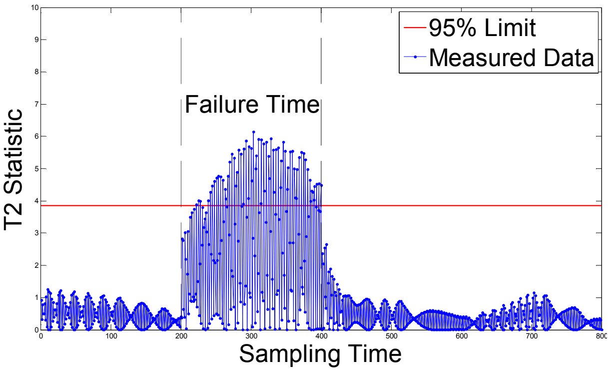

Figure 4. Fault detection result of wind power generation system based on PCA

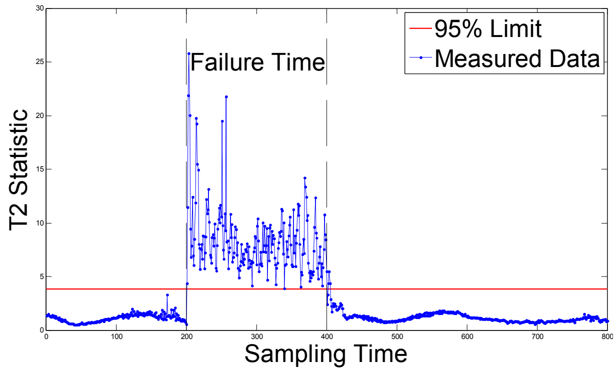

Figure 5. Fault detection result of wind power generation system based on NPCA

TABLE II. TWO KIND OF FAULT DETECTION METHODS PERFORMANCE COMPARISON

Multivariate statistical fault detection method	False alarm rate (%)	Missing alarm rate (%)
PCA	0	16.38
NPCA	0.37	0.13

VI. CONCLUSION

PCA method is usually suitable for fault detection under the steady conditions, because data of the steady conditions conforms to normal distribution. But when system is in the non-steady conditions, there will be a large number of phenomena of false alarm and missing alarm, and the main reason is the sample data of non-steady conditions could not meet normal distribution. As long as the data satisfies normal distribution, the T^2 control limit can detect fault effectively. The NPCA model is put forward under the periodic non-steady conditions can solve this problem. This model includes three steps: data normalization, PCA and fault detection based on T^2 control limit. Longitudinal standardization is the main part of data normalization, because it could make the transformed data meet the standard normal distribution to ensure fault detection valid through T^2 control limit. The NPCA model could not only detect fault data effectively, but also could reduce false alarm greatly, which improve the effectiveness of monitor tremendously.

ACKNOWLEDGMENT

We were grateful to the anonymous reviewers of a draft of this paper for their insightful remarks and useful comments.

REFERENCES

[1] J. F. Martins, V. Fernao Pires, A. J. Pires. PCA-Based On-Line Diagnosis of Induction Motor Stator Fault Feed by PWM Inverter. IEEE ISIE 2006, 2006, July 9-12.

[2] Gallagher N B, Wise B M Bulter, White S W, et al. Statistical process control tools for a semiconductor each process: improving robustness through model updating[C]. Proceeding of the ADCHEM, Banff, Canada,1997: 78-83.

[3] Wold S. Exponentially weighted moving principal component analysis and projection to latent stuctures. chemometrics and intelligent laboratory systems[C]. Proceedings of the 3rd Scandinavian Symposium on Chemometrics, 1994: 149-161.

[4] Weihua Li, Henry Yue H, Sergio Valle Cervantes,et al. Recursive PCA for adaptive process monitoring[J]. Journal of Process monitoring, 2000, 10(5): 471-486.

[5] Lane S, Martin E B, Morrios A J et al. Application of exponentially weighted principal component analysis for the monitoring of a polymer film manufacturing process[J]. Transactions of the institute of measure-ment and Control, 2003,25(1):17-35.

[6] T.Z. Wang, D.J. Gao, P. Liu , T.H. Tang. A Fault Detection Model Based on Dynamic Limit under Non-periodic Non-steady Conditions, J. Shanghai Jiaotong Univ.(Sci.), Vol.46, No.4, 2012, pp. 607-612.

[7] H. Polinder, F.F.A. van der Pijl, G.-J de Vilder, P. Tavner. Comparison of Direct-Drive and Geared Generator Concepts for Wind Turbines, IEEE International Conference on Electric Machines and Drives, 2005, pp.543 - 550.

[8] Wenying Chen. Application of Multi-scale Principal Component Analysis and SVM to the Motor Fault Diagnosis. 2009 International Forum on Information Technology and Applications, 2009.

[9] KARIM SALAHSHOOR, HESAM KOMARI ALAEI, IMAN NASR ISFAHANI. A New PCA Adaptive Rough Fuzzy Cluster Based Granulation Algorithm for Fault Detection and Diagnosis. 2010 IEEE International Conference on Control Applications, 2010, Sep. 8-10.

[10] Xiangyu He, Qinghua He. Application of PCA method and FCM clustering to the fault diagnosis of excavator's hydraulic system. Proceeding of the IEEE International Conference on Automation and Logistics, 2007, August 18-21.

[11] Shaohui Fan. Research on Extended Methods of PCA (Master thesis), Henan University, 2005.

[12] Richard A.Johnson,Dean W.Wichern. Applied Multivariate Statistical Analysis (6th Edition)(M).London:Prentice Hall, 2008.149-198

[13] R. Esmaili, L. Xu, D.K. Nichols. A New Control Method of Permanent Magnet Generator for Maximum Power Tracking in Wind Turbine Application. Proc.2005 IEEE Power Engineering Society General Meeting, vol.3, 2005, pp.2090-2095.

[14] X.Y. Zhang, S. He, X.B. Zhang. Study of Fault Diagnosis Wind Turbine Generator System[J], Journal of Xinjiang University, Vol.2, 2009.