Study on Complete Analysis of LRE Test Samples Based on PCA

Min WANG(王 珉), Niao-qing HU(胡鸟庆)
(*School of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha* 410073, *China*)

Abstract – **Incomplete data samples have a serious impact on the effectiveness of data mining. Aiming at the LRE historical test samples, based on correlation analysis of condition parameter, this paper introduced principle component analysis (PCA) and proposed a complete analysis method based on PCA for incomplete samples. At first, the covariance matrix of complete data set was calculated; Then, according to corresponding eigenvalues which were in descending, a principle matrix composed of eigen- vectors of covariance matrix was made; Finally, the vacant data was estimated based on the principle matrix and the known data. Compared with traditional method validated the method proposed in this paper has a better effect on complete test samples. An application example shows that the method suggested in this paper can update the value in use of historical test data.**

Key words – ***test sample***; ***data mining***; ***correlation analysis***; ***PCA***; ***complete analysis***

Manuscript Number: 1674-8042(2011)03-0217-05

doi: 10.3969/j.issn.1674-8042.2011.03.004

It is significant demanding for perfect fault mode knowledge to process Integrated System Health Management (ISHM) on ground test of Liquid Rocket Engine (LRE). There is no doubt that the test samples saved after ground testing are important information source for acquisition of fault knowledge[1]. But, incompletion is a common phenomenon in historical test data mining and has a serious impact on the effectiveness of data mining. The incompletion of test samples is the vacancy of part test data in the test samples, which is caused by the reasons as follows[2]:

1) Failure of transmission agent and recorder;

2) Some data being deleted because they are inconsistent with others;

3) Some data being not recorded because they can not be comprehended;

4) Not registering or data changing;

5) Some data being damaged in database.

The time-variant condition information of engine and test bed in the process of ground test is recorded in test sample, in which the test data of condition parameters constitute a time sequence that describes the variation of the parameters. There is on doubt that the vacancy of part data will make the condition information incomplete and affect data mining through lowering the start point of knowledge discovery and the accuracy of knowledge[3].

The cost of ground test of LRE is so high that the test can't be conducted frequently and any test is strictly forbidden. So, the historical test samples are so precious that they can't be abandoned in data mining just because of the vacant. It is an important preprocessing work to do complete analysis for incomplete samples to estimate vacant data[3].

This paper completed analysis of incomplete test samples by the statistical method. Based on the study of vacant data estimating and correlation analysis of condition parameters, Primary Component Analysis (PCA) was introduced and a complete analysis method for incomplete test samples was proposed, in which vacant data was estimated based on known data in test samples. The method in this paper was validated by an application example.

1 Complete analysis by the statistical method

Now, the complete analysis for incomplete sample is mainly done by three methods[4-6]:

1) Handling vacant data with machine studying;

2) Handling vacant data based on indiscernible relation of Rough Set;

3) Estimating vacant data based on known data in the samples by the statistical method.

For the statistical method doesn't need exact mathematic model, the vacant data is estimated with known data in the samples based on correlation analysis of condition parameters. The algorithms of complete analysis by the statistical method are[5,6]:

1) Average displacement: using the mean value of known data to replace the vacant data;

2) Cold Deck conjecture: using the data saved in the past studies to replace the vacant data;

3) Regression conjecture: estimating the vacant data by analyzing the relation of parameters based on regression analysis.

* **Received**: 2011-03-12
Project supported: This work was supported by National Natural Science Foundation of China(No. 51075391).
Corresponding author: Min WANG (wm198063@yahoo.com.cn)

The easiest and most common way to handle vacant data by statistical method was average displacement, which was comprehensively used in business and industrial production[4,5]. Average displacement algorithm is first used to estimate vacant data. A test sample recorded the condition of test-bed propellant filling system in a ground test is given in Tab. 1, which consists of twenty continual sample points of five key condition parameters in the stage of steady test[1]. Some data assumed is to be vacant and made in bold type.

Tab. 1 Test sample of test-bed in the stage of steady test

Sample point \ Parameter	Pejr (MPa)	Pxr (MPa)	PGr (MPa)	Gr (kg/s)	Pohr (MPa)
1	**1.741 3**	0.318 1	0.306 1	3.263 1	0.282 9
2	1.754 9	0.323 2	0.311 1	3.274 1	0.287 3
3	1.792 1	0.317 2	**0.305 2**	3.251 9	0.281 2
4	1.757 3	0.312 3	0.299 9	3.256 9	0.276 6
5	1.743 6	0.316 3	0.303 9	3.266 9	0.281 6
6	1.753 7	**0.321 6**	0.309 3	3.253 9	0.287 5
7	1.791 4	0.316 7	0.304 1	**3.249 8**	0.281 5
8	1.755 6	0.312 1	0.299 5	3.273 6	0.278 2
9	1.745 8	0.315 9	0.303 3	3.271 5	0.282 6
10	1.753 5	0.320 6	0.307 8	3.254 1	0.287 3
11	1.789 8	0.316 2	0.303 2	3.257 1	**0.280 6**
12	1.751 6	0.312 3	0.299 3	3.259 9	0.278 3
13	1.748 1	0.316 3	0.303 4	3.244 8	0.282 8
14	1.757 2	0.319 8	0.306 8	3.257 1	**0.285 7**
15	**1.784 1**	0.314 8	0.301 7	3.265 9	0.281 2
16	1.749 1	0.312 8	0.299 6	3.264 8	0.278 7
17	1.750 7	0.316 7	0.303 5	**3.265 5**	0.282 4
18	1.755 6	0.319 1	0.305 8	3.251 9	0.284 1
19	1.772 7	0.315 3	0.301 9	3.247 7	0.280 6
20	1.749 9	0.313 6	0.300 2	3.267 6	0.278 4

Estimate the vacant data by the average displacement algorithm and the estimated result is given in Tab. 2.

Tab. 2 Estimated result of average displacement algorithm

Sample point \ Parameter	Pejr (MPa)	Pxr (MPa)	PGr (MPa)	Gr (kg/s)	Pohr (MPa)
1	**1.754 4**	0.318 1	0.306 1	3.2631	0.2829
3	1.792 1	0.317 2	**0.301 1**	3.251 9	0.281 2
6	1.753 7	**0.316 8**	0.309 3	3.253 9	0.287 5
7	1.791 4	0.316 7	0.304 1	**3.248 3**	0.281 5
11	1.789 8	0.316 2	0.303 2	3.257 1	**0.277 8**
14	1.757 2	0.319 8	0.306 8	3.257 1	**0.277 8**
15	**1.754 4**	0.314 8	0.301 7	3.265 9	0.281 2
17	1.750 7	0.316 7	0.303 5	**3.248 3**	0.282 4

The estimated accuracy is usually weighed by the relative error between the estimated value and real value. In this paper, the aggregate error of vacant data δ is used to weigh the accuracy of complete analysis for the incomplete samples, δ is defined as

$$\delta = \sum_{i=1}^{N} \left| \frac{x_i - \hat{x}}{x_i} \right|, \quad (1)$$

where N is the number of vacant data, x_i is the real value, $\hat{x}_i$ is the estimated value.

For the above example, the aggregate error of complete analysis by the average displacement algorithm is $\delta = 0.095\ 9$.

With average displacement algorithm, the central value of test signal amplitude oscillating is calculated and used as estimated value of vacant data. As to the single parameter in the test sample, the estimated values of all vacant data are the same, which is not in disagreement with the practical situation. In the process of estimating vacant data, only the time sequence correlation of single parameter and the correlation information among the parameters is ignored, so the estimated result doesn't reflect the interdependence of all condition parameters and can't satisfy the need of the complete analysis.

2 Correlation analysis of condition parameters

During ground test, the variant of the condition parameters is not independent. In normal condition, there is higher statistical correlation between the test values of every parameter in different time, which is determined by some basic laws (such as mass conservation, energy conservation, etc)[7] and shows the range at which one parameter can change while the other parameters are given values.

The statistical correlation of the parameters can be described by correlation coefficient. As to two parameters, x and y, the correlation coefficient is defined as follows

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{D(x)D(y)}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \quad (2)$$

where $\bar{x}$ and $\bar{y}$ are the mean values of parameters' time series, N is the number of the sample points.

The range of the correlation coefficient is $-1 \sim 1$. If $r>0$, x and y are said to be positive correlation; if $r<0$, x and y are said to be negative correlation; and if $r=0$, x and y are said to be no correlation. Obviously, the larger $|r|$ is, the higher the statistical correlation between the two parameters is.

For Tab. 1, the result of correlation analysis is given in Tab. 3, which shows that there is statistical correlation at a certain degree between each two condition parameters of the propellant filling system.

The correlation coefficient reflects the correlation degree of some condition attributes of the system fsom the view of statistics. From Eq. (2), it can be seen that the correlation coefficient is in direct proportion to covariance , so covariance is the main carrier of the statistical correlation information[8].

Primary Component Analysis (PCA) constructs a lower dimensional space that intently reflects statistical correlation information included in the higher dimensional space through calculating covariance matrix of multiple parameters, based on which, tie regularities of data distribution are analyzed by translating higher dimensional space into lower dimensional space with PCA, and vacant data is estimated based on acquired information. Therefore, a method for complete analysis ofr incomplete test samples based on PCA was proposed in this paper.

Tab. 3 Result of Correlation Analysis

	Pejr	Pxr	PGr	Gr	Pohr
Pejr	1	−0.032 3	−0.032 4	−0.374 1	−0.133 6
Pxr	−0.032 3	1	0.991 9	−0.151 0	0.962 5
PGr	−0.032 4	0.991 9	1	−0.122 5	0.943 2
Gr	−0.374 1	−0.151 0	−0.122 5	1	−0.134 0
Pohr	−0.133 6	0.962 5	0.943 2	−0.134 0	1

3 Complete analysis based on PCA

3.1 Basic principle of PCA

In PCA, the data sample is written as an $m \times n$ matrix $X_{m\times n}$, m is the number of the parameters and n is the number of the sample points. Some integrated variables (the number of them is less than m) are constructed to intently reflect statistical correlation information included in m parameters. These integrated variables, named primary components, are calculated with the covariance matrix of $X_{m\times n}$ and they are mutually independent.

Suppose $x_i \in R^n$ is a test vector composed of n sample points of the parameter i, $X = [x_1, x_2 \cdots, x_m]$ is a test matrix composed of m parameters, the linear transformation of X is

$$\begin{cases} y_1 = l_{11}x_1 + l_{21}x_2 + \cdots + l_{m1}x_m = L_1 X, \\ y_m = L_{1m}x_1 + l_{2m}x_2 + \cdots + l_{mm}x_m = L_m X, \end{cases} \quad (3)$$

the constraint condition is

$$L_1^T \cdot L_i = 1,\ i = 1,2,\cdots,m. \quad (4)$$

Suppose a variable y_1 need to estimate X, which demands that y_1 contains enough information of X. According to classical information theory, the larger the covariance $\mathrm{var}(y_1)$, the more the information included in y_1. So y_1 can be calculated by finding L_1 under constraint condition (4) to make $\mathrm{var}(y_1)$ maximum, and y_1 is called the first primary component. If the information included in y_1 can't sufficiently represent X, the second primary component, y_2, can be added, in which L_2 is not only demanded to satisfy Eq. (4) and make $\mathrm{var}(y_2)$ maximum, but also needed to satisfy constraint condition (5):

$$\mathrm{cov}(y_1, y_2) = 0. \quad (5)$$

Similarly, p primary component of X can be defined, where X is divided into two parts

$$X = \hat{X} + \tilde{X}, \quad (6)$$

where $\hat{X}$ is named primary component matrix of X and $\tilde{X}$ is named residual matrix of X.

Suppose R is the covariance matrix of X, and $P \in R^{m\times m}$ is the unit eigen vector matrix of R. Obviously P is orthogonal matrix, that is $PP^T = I$, $PP^{-1} = I$, with which the theorem 1 is obtained.

Theorem 1: Make $\hat{P} \in R^{m\times p}$ and $\tilde{P} \in R^{m\times(m-p)}$, where the column vectors of $\hat{P}$ are the p eigen vectors of P corresponding to the max p eigenvalues, the column vectors of $\tilde{P}$ are the remained $m - p$ eigen vectors of P, X can be divided as follows[9]

$$X = \hat{X} + \tilde{X} = \hat{T}\hat{P}^T + \tilde{T}\tilde{P}^T, \quad (7)$$

where $\hat{T} \in R^{n\times p}$ and $\tilde{T} \in R^{n\times(n-p)}$ are called primary component score matrix and residual score matrix respectively. PThe above-mentioned is the process of modeling based on PCA, after which, condition parameter x can also be divided into two parts

$$x = \hat{x} + \tilde{x}, \quad (8)$$

where $\hat{x} = \hat{P}\hat{P}^T x$ is the projection of x in primary component space, $\tilde{x} = \tilde{P}\tilde{P}^T x$ is the projection of x in residual space.

3.2 Estimating vacant data with PCA

Suppose $e_i = [x_{1i}, x_{2i}, \cdots, x_{mi}]^T$ is the test vector of X at the sample point i, with Eq. (8), e_i can be divided into two parts: $e_{i1} = [y_{1i}, y_{2i}, \cdots, y_{ki}]^T$ and $e_{i2} = [y_{(k+1)i}, \cdots, y_{mi}]^T$, which satisfy

$$\begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix} = \hat{P}\hat{P}^T e_I. \quad (9)$$

Divide $\hat{P}$ into two matrixes T_1 and T_2 by line, where $T_1 \in R^{k\times m}$ and $T_2 \in R^{(m-k)\times m}$, then

$$\begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix} = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \hat{P}^T e_i. \quad (10)$$

Theorem 2: With matrix $A \in C^{m\times n}$, unitary matrixes $U = [u_1, u_2, \cdots, u_m]$ and $V = [v_1, v_2, \cdots, v_n]$ are obtained, which make Eq. (11) rational[10]:

$$U^H A V = \begin{bmatrix} D \\ 0_{(m-l)\times n} \end{bmatrix} \quad m > n,$$
$$U^H A V = [D \quad 0_{m\times(n-1)}] \quad m < n, \quad (11)$$

where $D = \mathrm{diag}[\sigma_1, \sigma_2, \cdots, \sigma_l]$ and $l = \min\{m, n\}$, suppose $m < n$ and $l = m$, then Eq. (11) can be written as

$$U^H A V = D[I_{m\times m}, 0_{m\times(n-m)}], \quad (12)$$

where I is unit matrix.

Definition 1: The generalized inverse matrix of $A \in C^{m\times n}$ is defined as[11]

$$A^+ = V^H \begin{bmatrix} I_{m\times m} \\ 0_{(n-m)\times m} \end{bmatrix} D^{-1} U. \quad (13)$$

Theorem 3: Suppose e_{i1} is the vacant data of X at the sample point i, e_{i2} is known, then e_{i1} can be estimated as

$$e_{i1} = T_1 T_2^+ e_{i2}, \quad (14)$$

where T_2^+ is the generalized inverse matrix of T_2.

Proof: From Eq. (10), it can be seen that $e_{i1} = T_1 \hat{P}^T e_i$ and $e_{i2} = T_2 \hat{P}^T e_i$, so $e_i = (\hat{P}^T)^{-1} T_2^+ e_{i2}$; from Eq. (14), it is obvious that $e_{i1} = T_1 \hat{P}^T e_i = T_1 \hat{P}^T (\hat{P}^T)^{-1} T_2^+ e_{i2} = T_1 T_2^+ e_{i2}$.

In the above study, the vacant data is supposed to submit the whole distribution of the test samples and is estimated with known data by calculating statistical characteristics of the test samples. When m is small, all eigenvectors of covariance matrix of test sample can be used to replace $\hat{P}$ with P.

Because of different dimensions of parameters, when calculating primary component based covariance matrix, the parameter with larger covariance may be focused on Ref. [12]. To reduce the effect of different dimensional parameters on statistical characteristics analysis of test samples, matrix X needs to be normalized by Eq. (15) before complete analysis

$$\bar{X} = [X - I_n U^T]/D_\sigma^{-1/2}, \quad (15)$$

where U and D are mean vector and variance matrix of X, respectively.

3.3 The algorithm of complete analysis based on PCA

Input: test sample matrix Y with vacant data,

Output: complete test sample matrix Y.

Step 1: Extract complete test vectors from Y with sample point to make complete matrix X;

Step 2: Normalize matrix X with Eq. (15), with the normalized matrix still signed as X;

Step 3: Calculate covariance matrix R and its eigenvalues and eigen vectors;

Step 4: Array eigen vectors according to corresponding eigenvalues in descending. If the number of parameters is larger, take the front p eigen vectors whose corresponding eigenvalue is larger and make matrix P;

Step 5: For the sample point which has vacant data, extract line vectors from P_s according to the serial number of parameters with vacant data in the test sample to make matrix T_1, and make matrix T_2 with the residual line vectors;

Step 6: Calculate generalized inverse matrix of T_2;

Step 7: Estimate vacant data based on Eq. (14).

3.4 Certification for the algorithm

Compile the proposed algorithm with MATLAB and estimate the vacant data in Tab. 1. The estimated result is given in Tab. 4 . The aggregate error of complete analysis is $\delta = 0.033\ 3$.

From Tab. 4, it can be seen that ,after complete analysis, the complete time series of the parameters reflect the whole distribution of test values, and are in agreement with the practical situation. The aggregate error of the proposed algorithm is less than that of average displacement algorithm, so the complete analysis method proposed in this paper is valid and has a better effect on complete test samples.

Tab. 4 Estimated result of the algorithm based on PCA proposedi in this paper

Sample point \ Parameter	Pejr (MPa)	Pxr (MPa)	PGr (MPa)	Gr (kg/s)	Pohr (MPa)
1	1.749 7	0.318 1	0.306 1	3.263 1	0.282 9
3	1.792 1	0.317 2	0.304 3	3.251 9	0.281 2
6	1.753 7	0.320 2	0.309 3	3.253 9	0.287 5
7	1.791 4	0.316 7	0.304 1	3.250 1	0.281 5
11	1.789 8	0.316 2	0.303 2	3.257 1	0.279 5
14	1.757 2	0.319 8	0.306 8	3.257 1	0.281 9
15	1.779 3	0.314 8	0.301 7	3.265 9	0.281 2
17	1.750 7	0.316 7	0.303 5	3.261 7	0.282 4

4 Application

The tangential vibration of the oxidant turbopump in TN618 ground test is given in Fig. 1, where there are short lived breakages in the transmission line of sensors near 91.5 s and 103.6 s, which make two segments of data recorded ineffective, as shown in imaginal line frame. Strictly speaking, in data mining, the phenomenon can't be deemed to the fault of sensors, because there are no fault symptoms of sensors. It should be attributed to incomplete test samples caused by failures of transmission agent and recorder. Before data mining, complete analysis is made by the proposed method.

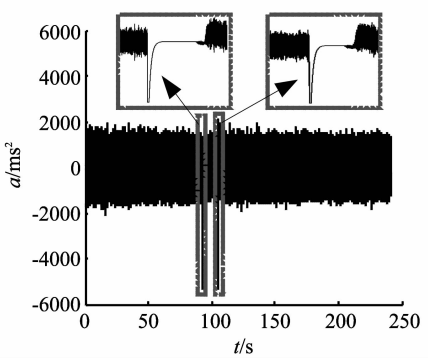

Fig. 1 Tangential vibration of the oxidant turbopump

In the ground test, the axial vibration and radial vibration of the oxidant turbopump are collected in synchronition with the tangential vibration, where the sampling frequency $f_s = 50$ kHz. Set the step as $M = 5\ 000$, the correlation coefficients of the tangential vibration with the axial vibration and the radial vibration are given in Fig. 2 and Fig. 3.

From Fig. 2 and Fig. 3, it can be seen that in the normal condition, there is statistical correlation at a certain degree of the three orientation vibrations of

the turbopump because of transmission and coupling. In this paper, the data segments of the three vibrations in 90 s～95 s are used to estimate the ineffective data segment near 91.5 s, and those in 102 s ～107 s are used to estimate the ineffective data segment near 103.6 s. The tangential vibration after complete analysis is shown in Fig. 4.

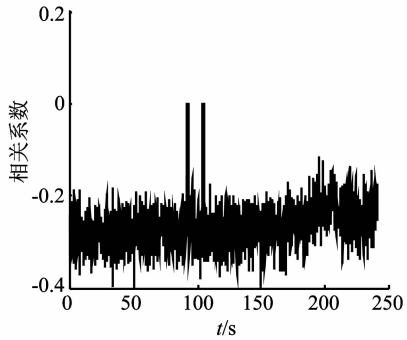

Fig. 2 Correlation coefficient of the tangential vibration with the axial vibration

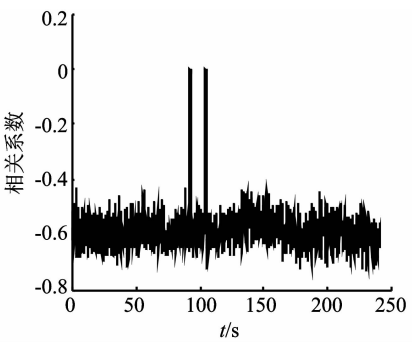

Fig. 3 Correlation coefficient of the tangential vibration with the radial vibration

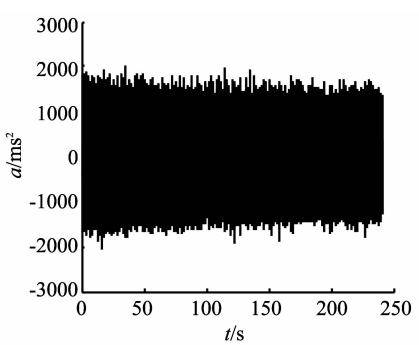

Fig. 4 Tangential vibration after complete analysis

As seen from Fig. 4, the estimated data submit to the whole distribution regularities of the test samples, reflects the practical tangential vibration information of the oxidant turbopump near 91.5 s and 103.6 s, reduces the effect of data ineffectiveness caused by the breakages on the transmission line of sensors, and improves the efficiency of data mining and accuracy of knowledge discovery.

5 Conclusions

The proposed complete analysis method based on PCA has many advantages (such as higher accuracy, less aggregate error, data submitting to the whole distribution regularities of test samples, etc) and is more effective than the traditional method (such as average displacement algorithm). But it needs to calculate covariance matrix of the test samples, and when there are lots of test data and more vacant data. So it is a key problem to advance the efficiency of calculation, which is very important point in subsequent studies.

References

[1] Min Wang, Niao-qing Hu, Si-feng Yang, et al, 2010. Study on application of fault simulation based fault-knowledge base. *Journal of Astronautic*, 31(4):1253-1258.

[2] Xiao-feng Guo, 1990. Liquid Rocket Engine Testing. Astronautics Press, Beijing, p. 102-103.

[3] Balaji Rajagopalan, Mark W Isken, 2001. Exploiting data preparation to enhance mining and knowledge discovery. IEEE transactions on systems. *Man and Cybernetics -Part C: Applications and Review*, 31(4): 54-61.

[4] Min Li, Yu Lu, Zhen-zhong Su, 2009. The methods of filling up vacancy in data preprocessing. 5(7):1546-1548.

[5] Soibelman L M, Hyunjoo Kim, 2002. Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, (1):39 -48.

[6] Somlinski, Walczak, Einax J W, 2002. Exploratory analysis of data sets with missing elements and outliers. *Chemosphere*, (49):233-245.

[7] Zhi Xiao, Yong Li, Chang-long Li, 2002. Method in data-preprocessing based on correlation analysis. *Journal of Chongqing University (Natural Science Edition)*, 25(6): 132-134.

[8] Vlad Popovici, Jean Philippe Thiran, 2002. PCA in autocorrelation space. *Proc. 16th International Conference on Pattern Recognition*, (2):132-135.

[9] Guo Li, Peng Zhang, Xue-ran Li, et al, 2008. Sensor fault detection based on dynamic principle component analysis. *Journal of Data Acquisition & Processing*, 23(3): 338-341.

[10] Yi Wu, Chao Li, Jian-shu Luo, et al, 2006. Fundamentals of applied mathematics. Higher Education Press, Beijing p. 77-78.

[11] Xiao-juan Chen, Wen-bin Guo, 2008. Application of singular value decomposition in g-inverses. *Journal of East China Normal University (National Science)*, (1): 25-29.

[12] Ting-feng Xue, Hong-gang Liu, Jian-jun Wu, 2007. Fault detection and diagnosis for sensors of LRE based on PCA. *Journal of Astronautics*, 28(6):1668-1672.