# Identification of the multivariable outliers using $T^2$ eclipse chart based on the improved Partial Least Squares regression

Liu Yunlian[1,a] Xi Yanhui[2,b] Liu Jianhua[3,c] Wu Tiebin[1,d*] Li Xinjun[1,e]

[1] Hunan University of Humanities, Science and Technology, Loudi, Hunan, 417000, China

[2] Electrical and Information Engineering College, Changsha University of Science & Technology, Changsha, Hunan 410077, China

[3]College of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou, Hunan 412007, China

[a]liuyunlian85@163.com, [b]804775693@qq.com, [c]jhliu0615@163.com ,

[d*]wutiebin81@163.com, [e]lixinjun80@163.com

* Corresponding author: Wu Tiebin

**Keywords**: Multivariable outliers; Partial Least Squares regression; $T^2$ eclipse chart

**Abstract.** When there is multi-variables in a sample, some samples which obviously disturb the relationships among variables are called outlier samples. However the presence of an extremely significant outlier sample tends to conceal some other outlier samples, which bringing great challenge to the identification of multivariable outliers. On this basis, a method of identifying the multivariable outliers in $T^2$ eclipse chart based on the improved Partial Least Squares regression (PLSR) is proposed. It is generally known that some outliers samples fail to be identified owing to significantly outlier samples are prone to influence the variance of $T^2$ chart. To solve this problem, a fuzzy variance computing method is put forward. The improved PLSR based $T^2$ chart can well overcome the masking effect in outliers identification.

## Introduction

The detection of multivariable outliers has been regarded a difficult problem. Although single variable is shown to be normal, some samples have found to apparently disturb the relationships among variables. Especially, as extremely significant outlier samples are presented, other outlier samples tend to be concealed, which leading to great difficulty in the detection of multivariable outliers.

Principal component analysis (PCA) and PLSR have been widely investigated and applied in the fault detection and the identification of outliers as they can extract the principal components of multivariable, reduce or eliminate the coupling among variables, and decrease the dimensions of variables [1]. PCA usually utilizes Q statistics and Hotelling $T^2$ statistic to monitor the outlier or fault in process data [2]. As PCA merely considers the features of independent variable in its application[3] and rarely concerns the association between independent and dependent variables, it is prone to show identifying mistakes in the detection of outlier and faults. S. Wold and C. Albano et al. proposed a PLSR method[4] . Such method not also integrates the ideal of PCA for extracting useful information in explanatory variables but also considers the explanatory effect of input on output of variables. It therefore can reflect the relationship between dependent and independent variables, and particularly is conducive to be used in the detection of the samples with multivariable outliers and faults. However, there is extremely significant outlier sample available; it also presents error in identification. To solve this problem, this research proposes a method of detecting the multivariable outliers in T$^2$ eclipse chart based on the improved PLSR.

**The identifying method of the multivariable outliers in T2 chart based on the improved PLSR**

It is assumed that $m$ th components are extracted from $n$ th samples using PLSR, the contribution ratio of $i$ ($i$=1,2,…, $n$) th sample to $h$ th component $t_h$ is $T_{hi}^2$ [5], we obtain

$$T_{hi}^2 = \frac{t_{hi}^2}{(n-1)s_h^2} \tag{1}$$

Where $s_h^2$ is the variance of $t_h$

Based on the equation (1), the contribution ratio of $i$ th sample to $t_1, \mathbf{L}, t_m$ is calculated as

$$T_i^2 = \frac{1}{(n-1)} \sum_{h=1}^{m} \frac{t_{hi}^2}{s_h^2} \tag{2}$$

Deviation tends to be produced in the analysis if $T_i^2$ is too large. The statistic proposed by Tracy et al. is usually used, as

$$\frac{n^2(n-m)}{m(n^2-1)} T_i^2 \sim F(m, n-m) \tag{3}$$

when

$$T_i^2 \geq \frac{m(n^2-1)}{n^2(n-m)} F_a(m, n-m) \tag{4}$$

, the $i$ th sample has large contribution ration, which may be a outlier. Where $a$ indicates significant level.

Based on Eqs.(2) and (4), it is obtained as

$$\sum_{h=1}^{m} \frac{t_{hi}^2}{s_h^2} \geq \frac{m(n^2-1)(n-1)}{n^2(n-m)} F_a(m, n-m) \tag{5}$$

$$c = \frac{m(n^2-1)(n-1)}{n^2(n-m)} F_a(m, n-m) \tag{6}$$

In the case of $m$ =2, we obtain

$$\frac{t_{1i}^2}{s_1^2} + \frac{t_{2i}^2}{s_2^2} = \frac{2(n^2-1)(n-1)}{n^2(n-2)} F_a(2, n-2) = c \tag{7}$$

As the T$^2$ ellipse defined in equation (7) , if all samples are shown to be within the ellipse, they are considered to be distributed uniformly without outlier points; otherwise, if the samples lie outside the ellipse or are near to the ellipse boundary, they are possibly outliers points; however the extremely significant outliers samples are likely to result in a sharp increase of the variances $s_1^2$

and $s_2{}^2$ in equation (7), which further making part of outlier samples being concealed. To deal with such defect, the computing method of variance is improved.

Assuming there is a data sequence $x=(x_1,\ x_2,\ \mathbf{L},\ x_n)$, the improved sample variance $s_I{}^2$ is written as

$$s_I{}^2=\frac{1}{n-1}\sum_{i=1}^{n} b_i(x_i-\bar{x})^2 \tag{8}$$

Where $\bar{x}$ average is value of the sequence $x$; $b_i$ is a fuzzy parameter. The further to the average value, the less the proportion of $b_i$ in the calculation of the variance, the equation (9) is apresented as

$$b_i=e^{-g} \tag{9}$$

Where $g$ is a coefficient, as shown in equation (10),

$$g=\begin{cases} 1 & if\ \ \frac{x_i-x_m}{s_d}\le 1 \\ t1 & if\ \ 1<\frac{x_i-x_m}{s_d}<k \\ t2 & if\ \ k\le\frac{x_i-x_m}{s_d} \end{cases} \tag{10}$$

Where $x_m$ denotes the median obtained by the arrangement of the data sequence $x$ in a increasing order; $s_d$ is standard deviation; $t1$ and $t2$ are coefficients ($t1\in(0,1]$, $t2\in(0,1]$ and $t1\le t2$), while $k$ is a parameter ($k\in[1,\ 2]$). The $T^2$ chart obtained by suing improved the equation of computing variance shows good anti-interference capability.

## The analysis of simulated results

The social economic indicator and electricity consumption of a county in Hunan province, China in 1990 to 2002 are listed in table 1[6]. The Electricity consumption values in 1995 and 2002 are shown to be outlier samples. However all samples are normal when using PCA to identify outliers, which indicating obvious error of PCA[6].

**Table 1 The social economic indicator and electricity consumption of a county in Hunan province, China in 1990 to 2002**

| Years | Social economic indicator | | | | Electricity consumption / kW.h |
|---|---|---|---|---|---|
| | Primary industry/ 10,000 yuan | Secondary industry/ 10,000 yuan | tertiary industry / 10,000 yuan | Per capita | |
| 1990 | 32221 | 14733 | 17229 | 948 | 12211 |
| 1991 | 33017 | 16161 | 18043 | 989 | 13985 |
| 1992 | 35414 | 17972 | 20660 | 1086 | 16024 |
| 1993 | 45265 | 24343 | 25076 | 1385 | 17622 |
| 1994 | 62507 | 31783 | 33364 | 1860 | 20407 |
| 1995 | 79166 | 43670 | 47671 | 2471 | 25216 |
| 1996 | 91329 | 56248 | 58407 | 2985 | 21377 |
| 1997 | 93813 | 70764 | 70601 | 3408 | 21276 |
| 1998 | 96279 | 79775 | 79289 | 3627 | 21778 |
| 1999 | 96184 | 84517 | 86434 | 3883 | 22558 |
| 2000 | 98461 | 87816 | 95667 | 4098 | 21979 |
| 2001 | 105418 | 101059 | 109202 | 5214 | 22774 |
| 2002 | 107900 | 114134 | 124721 | 5705 | 31607 |

$T^2$ ellipse chart is demonstrated in the figure 1 when using common $T^2$ ellipse chart to detect outliers. Since the 13th sample point in 2002 lies is found in the outside area of the $T^2$ ellipse, it is a outlier point; while the 6th point within the $T^2$ ellipse fails to be detected, which showing that the 13th point exerts a certain masking effect on the 6th point.

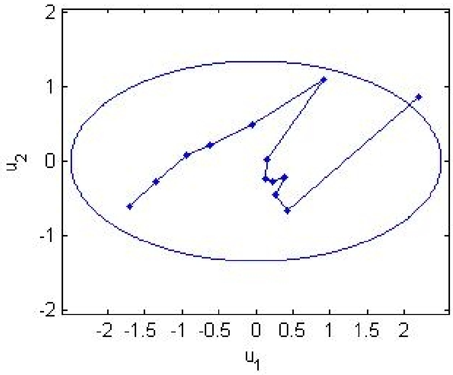

Figure 1 $T^2$ ellipse chart

The outliers detected in the improved $T^2$ chart are illustrated in figure 2.

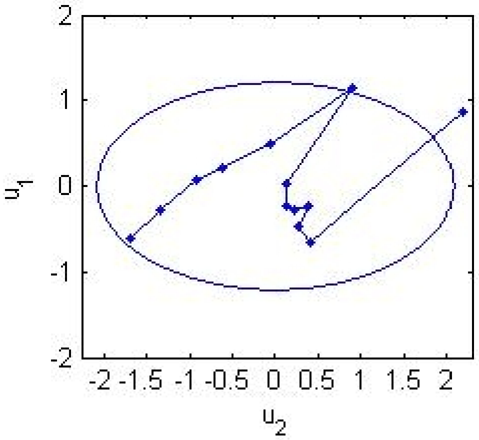

Figure 2 The improved $T^2$ ellipse chart

As shown in the figure 2, the improve method can recognize the 6th and 13th outlier points. First sample point which is close to the edge of the ellipse, is a mutation on power laod. Results indicate that the improved $T^2$ chart can well detect outliers and deal with the marking effect in the identification of outliers.

## Conclusions

In the case of the samples comprising multivariable, there is no apparently anomaly being found in single variable containing in the sample. Some samples which disturb the relationships among variables are considered as outlier samples. When there is extremely significant outlier sample available in the samples, the variance of the improved PLSR based $T^2$ chart can be influenced, which leads to the failure in the detection of some outlier samples. For solving this problem, this work put forwards a fuzzy variance computing method. The improved PLSR based $T^2$ chart is able to overcome the marking effect in the identifying multivariable outliers.

## Acknowledgements

This work was partially supported by The project supported by National Natural Science Foundation of China (NO. 61503131, NO. 61503131), and science and Technology Department of Loudi city, and Scientific Research Fund of Hunan Provincial Education Department(NO.14B097, NO. 15C0722)

## References

[1] Nie Yan Fang PCA and improved. The anomaly detection based on nearest neighbor rule. Computer engineering and design, 2008,29 (10):2502-2503.

[2] Zhang Xinrong, Xiong Weili, Xu Baoguo. Fault detection algorithm based on Q statistics[J]. Computer and applied chemistry, 2008, 25 (12): 1537-1542.

[3] Zhao Xiaoqiang; Wang Xinming, wangyingxiang. Based on PCA and KPCA TE process fault detection application research [J]. Automation and instrumentation, 2011, 32 (1): 8-12.

[4] Wold H． Partial Least Squares in Encyclopedis of Statistical Sciences [M ].New York：JohnWiley&Ston, 1985．

[5] Wang Huiwen. Partial least squares regression method and its application [M]. Beijing, National Defense Industry Press, 1999

[6] Mao Li Fan. Research on the technology of long-term load forecasting in power network planning [D]. Changsha: Hunan University, 2011