Screening of Abnormal Values Based on Centrifugal Rate

Xiaoli Zhang
College of Mathematics and Systems Sciences, Shandong University of Science and Technology, Shandong, Qingdao, 266590, China
Corresponding author’s e-mail: 276812529@qq.com

Abstract. In this paper, an outlier screening method based on centrifugal rate is proposed to screen outliers in data, which is to use the size of centrifugal rate to determine whether the data is significantly larger or smaller than other normal values of centrifugal rate. Taking the mean of data as the center, the distance between any point and the mean as the long half axis of the ellipse, and the distance between the normal dimension as the center, the model is established. The eccentricity based on the mean can be obtained, and so on, the eccentricity of all data based on the mean can be obtained. Finally, the centrifugal rate which deviates from the average centrifugal rate obviously surpasses other data is listed as the abnormal value, and the abnormal value is eliminated.

1. Introduction
Abnormal values are common in artificial errors, accidental errors and sampling errors. Sampling errors include some characteristics of data itself and some technical problems in sampling process; human errors include concealment, false reporting, misreading, misremembering, data input errors and data loss; accidental errors include inadequate acquisition equipment, confusion of measuring units, sudden instrumental indication, etc. Accidental errors include inadequate acquisition equipment, confusion of measurement units, sudden jump of instrument indication, sudden vibration, misoperation and other reasons that should not occur. If there are outliers in the data, the statistical analysis error will increase, so it is important to check the outliers.

Since Bernoulli defined outliers, there have been many studies on the test of outliers at home and abroad. Beckmen et al. [1] proposed two commonly used definitions of outliers: one is to interpret outliers as extreme values in a hypothetical distribution; the other is to treat outliers as impurity points, not from the same distribution as the main body of the data set. There are two kinds of outlier test methods: one is based on non-model, the other is based on model. Many scholars have studied the method based on model, such as Jin Libin [7] and so on. For the first-order spatial autoregressive model, according to the mean sliding model and variance weighted model, the score test statistics of outliers and their approximate distribution are given. Tian Yuzhu [5] and others put forward absolute value statistics, square statistics and adjusted square statistics for testing outliers in EXPAR model. The research shows that the above statistics are very effective for testing outliers. Rahmatullah et al. [3] In view of the fact that standardized Pearson residuals are only suitable for testing single outliers, a method for testing multiple outliers in multivariate logistic regression model is proposed by deleting group residuals. Kitagawa [2] proposes a model based on AIC criterion to test single outlier and multiple maximum (minimum) outliers, which overcomes the limitations of Dixon test and Grubbs test. Wang Zhijian [6] and others proposed to replace $\sigma$ with absolute deviation mean in order to improve the robustness of $\sigma$ in the test statistics of IO outliers. Data simulation shows that the improved
method has significantly improved the detection ability. Lv Qingzhe et al [4] used likelihood ratio method and Gibbs sampling method to detect AO outliers of ARMA model. The experiment shows that the latter method is better than the former. For non-model-based methods, many scholars have put forward many methods, such as: Zhang Deran [12] based on the three cases of abnormal large value, abnormal small value, both abnormal large value and abnormal small value, proposed the method of using jump degree of each point to test abnormal value, and gave the test statistics and their distribution in the form of theorem for the exponential distribution. Lalitha [10] and others proposed $Z_k$ statistics for exponential samples. Compared with Dixon statistics $D_k$, $T_k$ and $L_k$ statistics, $Z_k$ statistics had smaller detection deviation, and its critical value could be calculated. Jabbari [9] et al. extended the statistics of abnormal values of exponential distribution to gamma distribution, and compared them with Dixon statistics. The results show that the former is better than the latter. Wu [8] proposed the least square method to determine the number of maximum (minimum) outliers under normal samples. This method is simple and easy, and can overcome the Masking effect and Wamping effect.

2. Common methods of outlier test

In statistics, Grubbs test and t test are usually used to detect abnormal values, but they have certain limitations. Firstly, the above methods are only suitable for testing data subject to normal distribution. Secondly, due to the shielding of data, it is difficult to avoid Masking effect (positive differentiation) and Wamping effect (positive differentiation). Laida criterion is to assume that a set of test data contains only random errors, calculate and process them to obtain standard deviations, and determine an interval according to a certain probability. It is considered that errors beyond this interval are not random errors but gross errors, and the data containing such errors should be eliminated. The principle and method of discriminant processing are limited to the data processing of samples with normal or approximate normal distribution. It is based on the premise that the number of measurements is large enough. When the number of measurements is large enough, it is not reliable to eliminate gross errors with the criterion. Therefore, in the case of fewer measurements, it is better not to choose this criterion. The Grubbs test and Laida criterion are mainly introduced below.

The statistics of the Grubbs test are $G = \frac{x_i - \bar{x}}{s}$, in which $\bar{x}$ and $s$ are the mean and standard deviation calculated with all n data. According to the pre-determined confidence level and the number of data, if the $G$ value is greater than $G(n,0.05)$ ($n$ is the number of data to be checked, usually $\alpha = 0.05$), then $x_i$ is considered to be a n abnormal value and should be removed; otherwise, it should be retained. The Grubbs test contains suspicious values when calculating $\bar{x}$ and $s$. The existence of suspicious values makes the difference between $x_i$ and $x$ values and $s$ larger, and then makes the statistical values smaller, thus making mistakes easily. When there is more than one abnormal value in a set of measured values, the probability of Masking effect (positive difference) produced by Grubbs test increases, so the effect of testing abnormal value may not be optimal.

Assuming that the measured value is measured with equal precision, the arithmetic average $\bar{X}$ and residual error $v_i = x_i - \bar{X}$ of $x_1, x_2 \ldots x_n$ are calculated independently, and the standard error $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2}$ is calculated according to Bessel formula. If the residual error $x_m$ of a measured value $u_m = x_i - \bar{X}$ is $1 \leq m \leq n$, which satisfies $|u_m| = |x_i - \bar{X}| > 3s$ then $u_m$ is considered to be a bad value with gross error value, it should be eliminated.

3. Model One

The data to be examined is sorted from small to large, assuming that a normal value is known in the data. Based on this, the model is established: starting from the first data, the radius is the difference between the abscissa of the first data and the normal data, the circle is the center of the circle (also the center of the ellipse), and the distance between two points is the long half axis of the ellipse. So the centrifugal rate of the first data can be calculated based on the normal value, and so on, the centrifugal
rate of all data can be calculated based on the normal data. Finally, the data deviating from the average centrifugal rate to a certain extent are listed as abnormal values, and the abnormal values are eliminated. The following figure takes the first data as the normal value.

Let’s suppose that there are $N$ data $x_i$ in the arrangement, and the value of $y_i$ is the ordinal number of the arrangement. Repeatedly calculate one ($i = 1, 2 \cdots N$), such as $b_2 = |x_2 - x_1|$, $a_2 = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$, elliptical centrifugal rate $e = \frac{c}{a}$ in the figure above, and $c^2 = a^2 - b^2$ substituted for $a, b$ to obtain $c_2 = y_2 - y_1$, so the centrifugal rate $e_2 = \frac{c_2}{a_2} = \frac{y_2 - y_1}{\sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}}$, so the eccentric rate of an ellipse composed of any point and the first data is: $e_i = \frac{y_i - y_1}{\sqrt{(y_i - y_1)^2 + (x_i - x_1)^2}}$

where $1 < i \leq N, 0 < e_i < 1$. The average centrifugal rate is:

$$\bar{e} = \frac{1}{N-1} \sum_{i=2}^{N} \frac{|y_i - y_1|}{\sqrt{(y_i - y_1)^2 + (x_i - x_1)^2}}$$ (3.1)

The formula for deviating from the average centrifugal rate is:

$$E_i = \frac{|e_i - \bar{e}|}{\bar{e}}$$ (3.2)

When $E_i$ is obviously larger than other values, or when $E_i$ is ranked as an outlier with the largest first $n$, When $e_i = 1$, it is necessary to determine whether the value is abnormal

4. Model Two
The basis for model building is centered on the mean. The distance between the mean abscissa and the abscissa of any arbitrary point is taken as the short semi-axis of the ellipse. The distance between the mean and any point is taken as the long semi-axis of the ellipse. This model is used to consider the eccentricity of the relative mean of all the data. Let the independent variable $x_1, x_2 \cdots x_N$, and the dependent variable $y_1, y_2 \cdots y_N$, whose mean coordinate is $(X, Y)$, establish the following model, centered on the mean point, $b = |X - x_i|$ is the radius for the circle, then $a = \sqrt{(Y - y_1)^2 + (X - x_1)^2}$ is the ellipse long half axis, the ellipse center coordinate is $(X, Y)$, the short half axis Make an ellipse for $b = X - x_i$.

When the data dimension is $m_N$ dimension, the center point is taken as the reference, the Euclidean distance between the center point and any data is the long half-axis of the hyperellipsoid, and the Euclidean distance of the $(m_N - 1)$ dimension is the short half-axis of the hyperellipsoid, so as to judge the degree of the divergence of $m_i$-dimension ($i = 1, 2 \cdots N$), and N is the number of data. When the data form is not in the form of coordinates, the model only calculates the average of all the data, and then adds the average value to the data and sorts them together. The mean ordinate is the sorted serial number. The model diagram is as follows:
Figure 2. takes all data as normal values.

Point A coordinate is the first of the data, and \((X, Y)\) is the mean coordinate. We can get 
\[
a = \sqrt{(Y - y_1)^2 + (X - x_1)^2}, \quad b = |X - x_1|, \quad c^2 = a^2 - b^2 = (Y - y_1)^2
\]
so: 
\[
c = \frac{a}{\sqrt{(Y - y_1)^2 + (X - x_1)^2}}
\]
where \(0 < e < 1\).

The average centrifugal rate is:
\[
e_\bar{X} = \frac{\sum_{i=2}^{N} |y_i - y_1|}{\sqrt{(Y - y_1)^2 + (X - x_1)^2}}
\]
(3.4)

The formula for deviating from the average centrifugal rate is as follows:
\[
E_i = \frac{|x_i - \bar{X}|}{e}
\]
(3.5)

When there is a data point and the abscissa of the mean coordinate is the same, it can not form an ellipse at this time, and can only form a circle. At this time, the eccentricity is 1. When there is a data point and the mean ordinate is the same, the eccentricity is 0, so these two situations are not enough to explain whether this data is an outlier or not, and additional judgment is required. When the eccentricity of all the data is found, when the eccentricity of most of the data is similar, and a few eccentricities are significantly larger than most of the data, it can be roughly determined that the point is an abnormal point. In the case of linear fitting, after the point is excluded, the line is re-fitting and evaluated. If the fitting effect is good, the abnormal value is excluded. If the effect is not good, continue the steps of model two. Know that the requirements are met.

If there are \(n\) data \(x_i\) and there is no corresponding \(y_i\), then add \(\bar{X}\) to the data from small to large, rank the average to the second, the average ordinate \(\bar{Y}\) is several, and the repeated data as one.

5. Numerical Experiments

First, we use the Grubbs test to determine which of the following examples are outliers. The electrical performance indicators of a certain wire and cable product are measured to be 10 data of 1.56, 2.09, 2.09, 2.23, 2.33, 2.42, 2.42, 2.56, 2.66, etc., and Grubbs test is used to judge whether there is abnormal data. its calculation process is as follows:
\[
\bar{X} = \frac{\sum_{i=1}^{10} x_i}{10} = 2.245 \quad s = \sqrt{\frac{1}{10-1} \sum_{i=1}^{10} (x_i - \bar{X})^2} = 0.313
\]
because: 
\[
|x_1 - \bar{X}| = 0.685 > |x_{10} - \bar{X}| = 0.415 \quad x_i = x_1 = 1.56
\]
\[
G_i = \frac{|x_i - \bar{X}|}{s} = \frac{0.685}{0.313} = 2.19
\]

Look up table 1. Grabbs critical value \(G_{(0.05,10)} = 2.18\), so \(2.19 = G_i \geq G_{(0.05,10)} = 2.18\), Therefore, the corresponding \(x_1 = 1.56\) is considered as abnormal data.
After taking out $x_1 = 1.56$, the total test data is $n - 1 = 9$. Repeat the above steps and calculate as follows:

$$\hat{X}^* = \frac{\sum_{i=1}^{9} x_i}{9} = 2.321 \quad s^* = \sqrt{\frac{1}{9-1}\sum_{i=1}^{9}(x_i - \hat{X}^*)^2} = 0.212$$

because: $|x_1 - \hat{X}^*| = 0.231 < |x_9 - \hat{X}^*| = 0.339$, so: $x_1 = x_9 = 2.66$

$$G_l = \frac{|x_l - \hat{X}^*|}{s^*} = \frac{0.339}{0.212} = 1.599$$

Looking up table 1, Grabbs critical value $G_{(0.05,9)} = 2.11$, so $1.599 = G_l < G_{(0.05,9)} = 2.11$, that the corresponding $x_9 = 2.66$ is not abnormal data. Therefore, all data values except $x_1 = 1.56$ are valid data values.

Next, model 2 is used to find outliers. First, the average $\bar{Y} = 2.245$ of all data is calculated. Then, the average value is added to the data from small to large, and the average is arranged to the nearest, which is the abscissa coordinate. The duplicated data are treated as one. Formulas (1) and formulas (2) in model 2 can be substituted for the following table:

| $X$ | 1   | 2x3 | 3   | 4   | 5   | 6x2 | 7   | 8   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $\bar{Y}$ | 1.56 | 2.09 | 2.23 | 2.245 | 2.33 | 2.42 | 2.56 | 2.66 |
| Centrifugal rate | 22.3% | 7.7% | 1.5% | mean | 8.5% | 8.7% | 10.4% | 10.3% |

As can be seen from the above table, the first data eccentricity is significantly larger than other data, and should be judged as an abnormal value. So it is feasible for the model two to find outliers that are different from other normal data.

Numerical experiments were performed in a linear fit using Model One. Knowing the line $y_w = x_w + 1$, we can get the line through (1,2), (2,3), (4,5), (6,7), and we assign a singular value of (3,7). The new line that can be newly added after adding the outliers to the matlab is:

$$y_v = 0.959459459459460x_v + 1.72972972972973$$

The centrifugal rate of the coordinates of (3,7) is the largest relative to the first data, and whether $E_i = \frac{|e_i - \bar{e}|}{\bar{e}}$ is obviously larger than other data. After entering the above data points in Matlab, $e = \text{abs}(y_1 / 2) / \sqrt{(y_1 - 2)^2 + (x_1 - 1)^2}$, where $x_i$ and $y_i$ are the corresponding coordinate values in Matlab. The centrifugal rate of coordinates (3,7) is 0.928476690885295. The centrifugal rate of other points is 0.7071067811865, and $E_2 = \frac{|e_2 - \bar{e}|}{\bar{e}} = 23.57\%$, so it can be concluded that the corresponding points (3,7) of centrifugal rate $e_2$ are abnormal values. It shows that this model is suitable for linear fitting.

If there are multiple outliers, see if the model can find the most divergent data. Or the above example is known as the straight line $y_w = x_w + 1$. After adding the (3,7) and (8,6) outliers through (1,2), (2,3), (4,5), (6,7), (7,8), (9,10) and (10,11), the new regression equation obtained by matlab can be obtained with the data containing the outliers:

$$y_{gb} = 0.817567567567568x_{gb} + 2.013513513513515$$

By substituting the data into the formula $e_i = \frac{c_i}{a_i} = \frac{\hat{Y} - y_i}{\sqrt{(\hat{Y} - y_i)^2 + (\hat{X} - \bar{X})^2}}$, we can get that the centrifugal rates of coordinates (3,7) and (8,6) are 0.1961, and those of other normal values are 0.7071, so $\bar{e} = 0.6049$, then $E_2 = \frac{|e_2 - \bar{e}|}{\bar{e}} = 31\%, E_6 = \frac{|e_6 - \bar{e}|}{\bar{e}} = 30\%$.

The other normal values deviate from the average eccentricity $E = 0.18\%$, so the abnormal value deviates from the average eccentricity much farther than the other normal values, that is, the two abnormal values can be excluded. This experiment indirectly verifies that the larger the eccentricity, the more the value is the abnormal value, and the other is that the data should be compared with all the data, and the data that is obviously different from the eccentricity of most of the data is listed as the abnormal value.
6. Summary
The existence of outliers has a serious impact on predicting future changes. This paper proposes an eccentricity-based outlier detection algorithm for linearly correlated samples in low-dimensional space. The algorithm can deviate from the mean eccentricity. Therefore, there is a way to refer to the elimination of outliers. If you do not eliminate the outliers and directly analyze the data, it will lead to mistakes in decision making, so it should be based on real and reliable data before making decisions. The method can also exclude data that is close to the mean value, without calculating their eccentricity, which can simplify the calculation amount, so that the abnormal data can be excavated more quickly.

References
[1] Beckman R J, Cook R D. Outlier [J]. Techonmetrics, 1983(25): 119-149.
[2] Kitagawa G. On the Use of AIC for the Detection of Outliers [J]. Technometrics. 1979, 21(2): 193-199.
[3] Rahmatullah A H M. Ali S H. Identification of Multiple Outliers in Logistic Regression [J]. Communications In Statistics Theory and Methods, 2008, 37(11): 1697-1709.
[4] Lu Qingzhe, Dialect Rong and Lin Jinguan. Comparison of two time series outlier mining methods [J]. Statistical research, 2005, 22(1): 51-54.
[5] Tian Yuzhu, Ran Yanping, Chen Ping. Diagnosis of abnormal points in EXPAR time series model based on extreme value theory [J]. Mathematical statistics and management. 2010, 29(04): 628-636.
[6] Wang Zhijian, Wang Binhui. An improved time series IO outlier detection method [J]. Statistics and decision-making. 2014, 30(22): 4-6.
[7] Jin Libin, Dai Xiaowen, Shilei. Abnormal Value Test of Mixed Spatial Autoregressive Model [J]. Statistics and Decision-making, 2015, (01) : 12-16.
[8] Wu J. On determining the number of upper or lower outliers in mormal sample by least squares procedure[J]. Journal of Information and Optimization Sciences. 2003, 24(1) : 61-91.
[9] Jabbari N M, Nasiri P Detecting Outliers in Gamma Distribution [J]. Communication in Statistics-Theory and Methods, 2010, 39(4) : 698 -706.
[10] Lalitha S, Kumar N. Multiple outlier test for upper outliers in an exponential sample [J]. Journal of Applied Statistics. 2012, 39(6) : 1323-1330.
[11] Wang Bingxing. Test of abnormal data in exponential distribution [J]. 2001, 8(3): 255-259.
[12] Zhang Deran. Test method of outliers in statistical data [J]. Statistical research, 2003, 20(5): 53-55.