The 5th National Symposium on Engineering Risk & Insurance Research
17-19 Aug 2018, Harbin, China
D.G. Lv, H.W. Huang (editors)

A simple outlier detection method for the multiple dimensional problem

Shuo Zheng[1], Qinxuan Deng[2] and Yuxin Zhu[3]

[1]*State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, No.8 Donghu South Road, Wuhan, PR China, Email: zhengshuo@whu.edu.cn.*
[2]*State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, No.8 Donghu South Road, Wuhan, PR China, Email: January@whu.edu.cn.*
[3]*State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, No.8 Donghu South Road, Wuhan, PR China, Email: yuxinzhu@whu.edu.cn.*

Abstract: Outlier is attached importance in statistics and engineering, because it might result in misleading identification results. Many outlier detection methods have been proposed such as least median of squares (LMS), minimum volume of ellipsoid (MVE) and principal component analysis (PCA) etc. However, the mathematical complexity of many proposed methods hampers their practical uses. This paper develops a method which is easy to understand, simple to implement, called resampling half-means with Mahalanobis distance (RHM-MD). It can be used to detect outlier in case of multiple dimensional and small sample problem considering the correlation between different variable characteristics. Not only the outlier row vector but also the outlying component in the vector can be detected. The proposed method is illustrated and validated through a set of five dimensional correlated data which is simulated with ten known outliers. It turns out that the proposed method can detect the outliers rationally. Finally, RHM-MD is applied to detect outliers in the real data provided by TC304 for the student contest.

Keywords: Outlier detection, half-means, Mahalanobis distance, multiple dimension, correlation.

1. Introduction

An outlying observation, or “outlier,” is one that appears to deviate markedly from other observations (Frank 1969). It can be detected by its abnormal performance, because it has some characteristics that are distinct from other surrounding data. There are many reasons leading to outliers, most of which are caused by measurement error, e.g., human error, instrument failure, or unknown environmental disturbances (Hawkins 1980; Barnett and Lewis 1994). It is also possible that such outliers do exist in real data. In practice, directly incorporating measurements or observations with outliers into data analysis might lead to significant bias. The purpose, therefore, of outlier detection is to find out the data patterns implied by a small amount of abnormal data that are significantly different from the conventional data patterns (Han and Kamber 2001), so as to eliminate them and reduce their impacts on statistical inferences (Rousseeuw 1991).

The traditional method for outlier detection includes least median of squares (LMS), minimum volume of ellipsoid (MVE) and principal component analysis (PCA) (Rousseeuw 1987). Rousseeuw extended the median methodology to a robust regression technique, i.e., LMS. The LMS is based on the minimum median of squares, because the median will be unaffected by outliers for a large number of observations. Such modifications have no effect on the sample median. Although LMS could find all unusual observations, but may erroneously identify normal observations as outliers (Fung 1993). MVE is a robust multivariate method for dealing with the problem caused by multiple outliers (Rousseeuw 1985). The main idea is to assume that all the sample data constitute a dimensional hyper-elliptic sphere. Picking at least half of the samples from all the sample data, and searching for the smallest volume of hyper-ellipsoids. The problem with the MVE method is that it requires expensive computation cost, especially in the case of multi-dimensional problems because of profound computational complexity of hyper-ellipsoid. There are also numerous methods for outlier detection developed based on the principal component analysis (PCA) of the covariance matrix (Mardia and Kent 1979).

The key idea of the outlier detection method based on matrix decomposition is to use principal component analysis to find those outliers that violate the correlation between data. Singular value decomposition (SVD) is a numerically stable method for principal component analysis (William and Stephen 1998). After that the genetic or evolutionary methods (Walczak 1995) and projection pursuit (Ammann 1993) are utilized for robust PCA/SVD, while the shortcoming is that it requires unreasonable computational efforts in high dimension problems (Woodruff and Rocke 1994).

While these methods are theoretically sound, the complexity of many proposed methods, in terms of both comprehension and implementation, hampers their practical uses. Considering the problem being explored is to find out the outliers from the site investigation dataset at a clay site, which is a multi-dimensional and small sample problem. William and Stephen (1998) proposed a method called resampling by half-means (RHM) to detect outliers by studying the distribution of observation vector lengths obtained by sampling without replacement from the original data set. RHM is easy to understand, simple to implement, and it can also tackle the difficulties under the assumption of independent variables. In this paper, RHM is extended to be feasible in multi-dimensional outlier detection problems, called RHM-MD, which considers correlation between various characteristics through calculating Mahalanobis distance. Row vectors with outliers in a data matrix **X** can be found, and then Euclidean distance which is calculated according to each de-trend and normalized column vector of each variable in the data matrix can reflect the contribution of every component to the occurrence of outlier.

This paper starts with development of the RHM-MD method, followed by illustration and validation of the proposed approach using a set of simulated data. Then, the RHM-MD is applied to solve the TC304 student contest question.

2. Methodology

2.1 Outlier detection method RHM-MD

The RHM-MD method consists of two simple concepts, Mahalanobis distance (MD) and resampling method, which are introduced briefly in this section and combined to develop the proposed method.

The Mahalanobis distance was proposed by the P. C. Mahalanobis (Mahalanobis 1936) and represented the covariance distance of the data. For a p-dimensional multivariate vector $\mathbf{X}_i$ (i = 1, 2, ..., n) the MD is defined as:

$$\mathbf{D}(i) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})} \quad \text{for } i = 1,..., \text{n} \qquad (1)$$

where $\mathbf{X}_i = [x_{i,1}, x_{i,2}, ..., x_{i,\mathrm{p}}]$is the i-th row vector in the data matrix **X**; $\mathbf{X}=[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_\mathrm{n}]^\mathbf{T}$ is a n×p data matrix consisting of n observations and p variables; $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the 1×p row vector which is the estimated arithmetic mean of each variable and the p×p sample covariance matrix containing the correlation between different variables, respectively. Eq. 1 provides an effective way to calculate the similarity of two unknown sample sets (Mark and Tunnell 1985). It takes into account the connection between various characteristics and is scale-invariant, which is independent of the measurement scale. The Mahalanobis distance gives a judgmental observation. The larger the distance is, the more likely the point is an outlier. However, if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are calculated according to entire data matrix **X** containing the outliers, the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ estimators will be affected by the outliers leading to misidentification of outliers. It is hence important to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on the regular data set to detect outliers, which are however unknown before analysis. Thus, in the RHM-MD method, 50% of the total set of data are determined to be normal data (William and Stephen 1998). The theoretical connotation is that half of the data must be normal, otherwise it is meaningless to search for outliers. For reasonably choosing the half of the data, a resampling method is adopted in this paper.

Resampling methods are used to generate a distribution of statistics of interest by repeatedly calculating these statistics from randomly selected subset of the data (Hartigan 1969). If the estimation results of each data subset are consistent or relatively close, the authenticity of the inferred results is more convincing since each sample subset is independent and the sampling method is the same. Outliers are detected by examining the distribution of MDs obtained from sampling without replacement from the original data set (Efron 1979).

Fig. 1 shows a flow chart for the implementation of RHM-MD. Initially, select the subset of data without replacement from the entire sample **X** until up to the size of 50% (i.e., n/2) of the **X**. These sampled observations are placed in a new matrix, **X**(*k*) (i.e., *k*=1, ..., N_b, where N_b = the number of resampling). Then the covariation matrix **Σ**(*k*) and the mean **μ**(*k*) of **X**(*k*) are calculated to determine the *k*-th center of data subset. Then, the MDs of entire data **X** are calculated based on the **μ**(*k*) and **Σ**(*k*) using Eq. 1. Store the **D**(*k*) and repeat the steps above until the prescribed number of resampling N_b is reached. **D** stores all the **D**(*k*), and finally a histogram of all Mahalanobis distances in **D** can be made to see distribution of MDs.

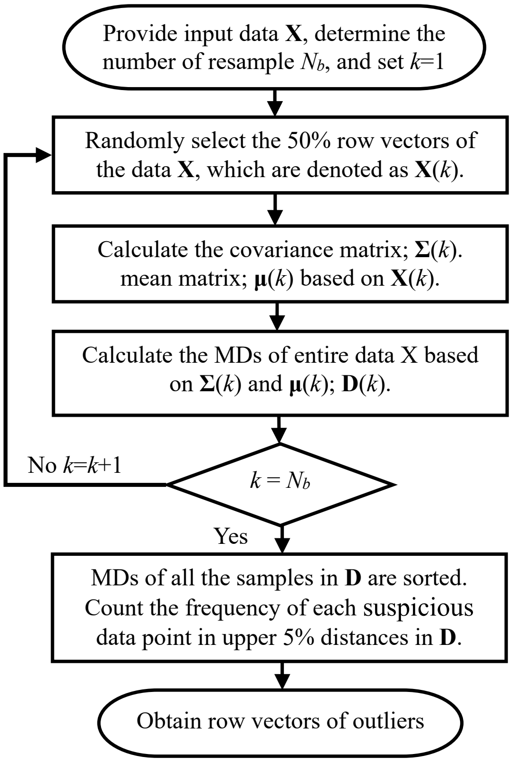
Provide input data X, determine the number of resample N_b, and set k=1
Randomly select the 50% row vectors of the data X, which are denoted as X(k).
Calculate the covariance matrix; Σ(k). mean matrix; μ(k) based on X(k).
Calculate the MDs of entire data X based on Σ(k) and μ(k); D(k).
No k=k+1
$k = N_b$
Yes
MDs of all the samples in D are sorted. Count the frequency of each suspicious data point in upper 5% distances in D.
Obtain row vectors of outliers

Figure 1. Flow chart of RHM-MD

2.2 Determination of outliers

For determining the rows of outliers in the data matrix **X**, the RHM-MD method inherently provides a simple way to identify outlying observations. Mahalanobis distances of all the samples in **D** are sorted. The upper 5% distances in **D** are considered as abnormal distances, which means that the corresponding data are suspicious data points. Then, the number of times each data appears in the upper 5% of the distance is counted. Considering that if there is no outlier, theoretically, the count of each data point occurring in the upper 5% of **D** the distance will be evenly distributed. If the number of resampling is relatively large, the percent of each point in the upper 5% of the distance will be approximately stable at 1/n (i.e., n is the number of data points). If the **X** contains one or more outliers, the percentage of outliers in the upper 5% of samples will be greater than 1/n. 1/n can be considered as a robust criterion for user to detect outliers.

When dealing with multidimensional data, the above work provides a simple way to determine row vectors of outliers $\mathbf{X}^{\mathbf{out}}_i$ and regular data $\mathbf{X}^{\mathbf{r}}_i$. To identify the exact outlier component $x^{\text{out}}_{i,j}$ (e.g., i = 1, 2, ..., n; j = 1, 2, ..., p) in the outlier row vector $\mathbf{X}^{\mathbf{out}}_i$, the Euclidean distance between each variable in $\mathbf{X}^{\mathbf{out}}_i$ and the center of corresponding variable, which is estimated using regular data, is calculated. The larger the distance represents that the component makes the greater contribution to the row of data to be an outlier $\mathbf{X}^{\mathbf{out}}_i$. In case of the impact of trend and scale of variables on outlier detection, the trend of the column vector $\mathbf{X}_{\cdot j}$ is removed and then residual vector $\boldsymbol{\varepsilon}_{\cdot j}= [\varepsilon_{1,j}, \varepsilon_{2,j}, ..., \varepsilon_{\text{n},j}]^{\text{T}}$ of the j-th column vector $\mathbf{X}_{\cdot j} = [x_{1,j}, x_{2,j}, ..., x_{\text{n},j}]^{\text{T}}$ is normalized. The Euclidean distance $\boldsymbol{d}^{\mathbf{out}}_{\cdot j}$ between each variable of normalized $\boldsymbol{\varepsilon}^{\mathbf{out}}_{\cdot j}$ and the center of corresponding variable normalized $\boldsymbol{\varepsilon}^{\mathbf{R}}_{\cdot j}$ is calculated to determine the most probable outlying variable $x^{\text{out}}_{i,j}$ in the outlier row vector.

3. Illustrative examples

For illustration and validation, the proposed method is applied to a set of simulated data which is simulated from a weighted mixture distribution of 5-dimensions normal distribution and triangular distribution in this section. Then, the proposed method is applied to detecting outliers in the dataset provided for TC304 student contest.

3.1 Simulated example

To generate the dataset that contains both regular data points and outliers, two sub-datasets are generated, namely the regular dataset D_r and the outliers D_f. In order to generate this set, a weighted mixture distribution of entire dataset D_t is used (Yuen and Mu 2012):

$$p(D_t) = (1-\rho)\cdot G(D_t|\mu,\Sigma) + \rho\cdot f(D_f) \tag{2}$$

where the multi-dimensional normal distribution $G(D_t| \ \mu, \Sigma)$ is used for the regular data points, and $f(D_f)$ is used for the outliers. In this simulated examples, $G(D_t| \ \mu, \ \Sigma)$ is a multi-dimensional normal distribution and it can be rewritten as:

$$G(D_t|\mu,R,\sigma)= \frac{1}{(2\pi\sigma^2)^{P/2}}\frac{1}{|R|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu)^T R^{-1}(\mathbf{x}-\mu)\right\} \tag{3}$$

where μ in Eqs. 2 and 3 is the mean of the P-dimensional random vector **X**; R is a P×P correlation matrix; σ is the standard deviation and it is taken as constant; $f(D_f)$ is a mixture of triangular distributions (Yuen and Mu 2012):

$$\begin{aligned} f(D_f) &= 0.5\times T(D_f\,|-5\sigma,-4\sigma,-3\sigma) \\ &\quad + 0.5\times T(D_f\,|\,3\sigma,4\sigma,5\sigma) \end{aligned} \tag{4}$$

where $T(D_f\,|\,l, m, r)$ is a triangular distribution:

$$T(D_f\,|\,l,m,r) = \begin{cases} 2(D_f - l)/(r-l)(m-l) & l \le D_f \le m \\ 2(r-D_f)/(r-l)(r-m) & m < D_f \le r \end{cases} \tag{5}$$

where $D_f = m$ is the mode value of the triangular distribution.

The variable ρ in Eq. 2 is the contaminated level parameter controlling the weighting of the contaminated distribution $f(D_f)$. This parameter can be interpreted as the probability of a data point being an outlier. It is worth noting that the value of ρ and the probability distributions in Eqs. 3 and 4 are assumed unknown in the identification process. They are only used to generate of data points. Note that the expected number of outliers N_0 in the entire dataset D_t is

Table 1. Description of simulated data

Dimension P	N	N_0	σ	ρ
5	30	10	1.5	0.33

$$N_0 = \rho\cdot N \tag{6}$$

where N is the number of entire dataset. Then, the proposed method is used to detect outliers in the entire dataset D_t. Table 1 summarizes the dimension of simulated data, the number of entire dataset N, the number of outlier dataset N_0 and the contaminated level parameter ρ in Eq. 6.

Specifically, the regular dataset is generated from a 5-dimensional normal distribution $G(D_t|$ $\mu, R, \sigma)$ with $\mu = [0,0,0,0,0]$, $\sigma = 1.5$, and

$$R = \begin{bmatrix} 1 & 0.82 & 0.67 & 0.55 & 0.45 \\ & 1 & 0.82 & 0.67 & 0.55 \\ & & 1 & 0.82 & 0.67 \\ & & & 1 & 0.82 \\ sym. & & & & 1 \end{bmatrix} \tag{7}$$

A total of 30 data points are simulated with the expected number of 10 outliers because the contaminated level 0.33 is imposed. Fig. 2 shows the results of the simulation run, where the 20 regular data points and 10 outliers are represented by squares and circles, respectively.

Fig. 3 displays MDs histogram of all the samples **D** from resampling procedure described in the section 2.1 and the number of resampling is $N_b = 5\times10^5$. The appearance of part B in Fig. 3 corresponds to extreme MDs caused by the suspicious outliers. The higher the extreme distance is, the more likely the points are outliers, which provides a sound visual diagnostic for determining which distance is selected as outliers that belong to another distribution. As shown in Table 2, there are sixteen suspicious row vectors occurring in the upper 5% MDs (i.e., the number of 5% MDs is $5\%\times30\times N_b=7.5\times10^4$, which is equal to the sum of the third column in Table 2). The frequency of each suspicious data point in upper 5% MDs in **D** indicates that nine of them (see bold results in Table 2) are greater than the criterion, i.e. 1/n = 0.033. In general, identified outliers index are nearly consistent with known outliers' index

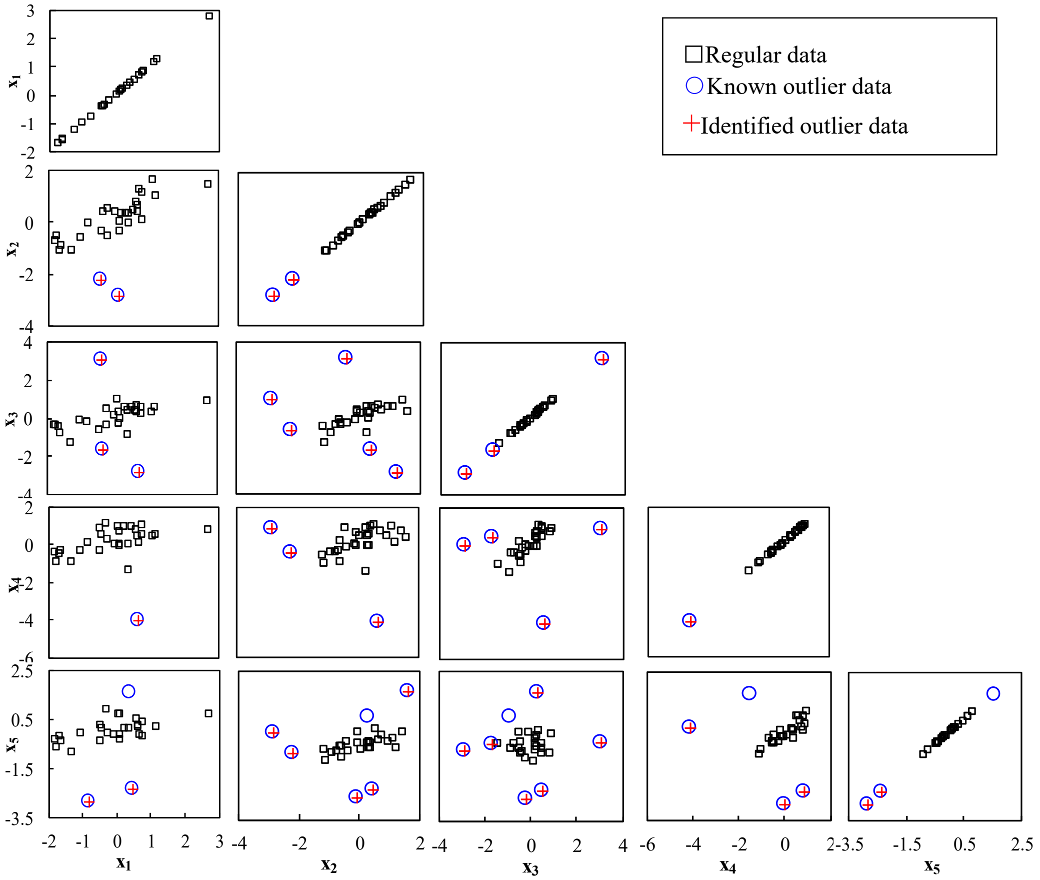

Figure 2. Simulated data and identified outlier data

except the 30-th row vector. However, the frequency of the 30-th vector, i.e., 0.0297, is closed to 0.033. In addition, the Euclidean distances $\boldsymbol{d}^{\mathbf{out}}_{\cdot,j}$ of identified outliers, which are calculated based on the procedure proposed in Section 2.2, are shown in Table 3. The most probable outlying variables with the largest distance, which are framed by black lines in Table 3, are shown in Fig. 2 by red crosses correspondingly. Compared with the known simulated outliers represented by blue circles in Fig. 2, there is only one outlying data point, which is not detected and it corresponds to the 30-th row vector in Table 2. This represents that the proposed method can detect the most probable outliers and $\boldsymbol{d}^{\mathbf{out}}_{i,\cdot}$ can effectively reflect the contribution of variables to the row of data to be an outlier $\mathbf{X}^{\mathbf{out}}_{i}$.

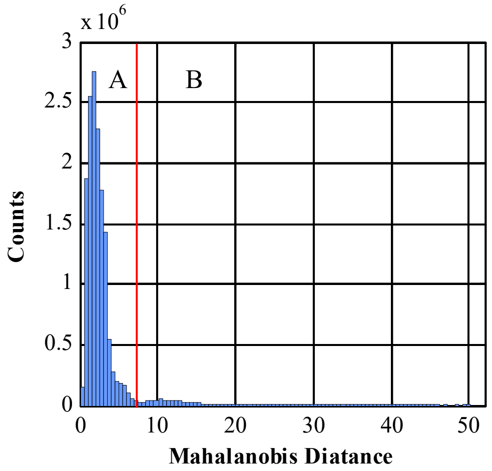

Figure 3. Distribution of Mahalanobis distances using RHM-MD with outliers

Table 2. RHM-MD results of the simulated data

Known outlier row index	Row index	Counts	Frequency (%)
	1	273	0.04
3	**3**	**65926**	**8.79**
4	**4**	**35918**	**4.79**
9	**9**	**28597**	**3.81**
12	**12**	**152418**	**20.32**
13	**13**	**32996**	**4.40**
14	**14**	**27160**	**3.62**
15	**15**	**249991**	**33.33**
	16	1116	0.15
19	**19**	**70182**	**9.36**
	20	723	0.10
21	**21**	**61603**	**8.21**
	25	210	0.03
	27	80	0.01
	28	416	0.06
30	30	22299	2.97

Table 3. Results of Euclidean distances for the simulated case

Identified outlier index	$d^{out}_{\cdot,1}$	$d^{out}_{\cdot,2}$	$d^{out}_{\cdot,3}$	$d^{out}_{\cdot,4}$	$d^{out}_{\cdot,5}$
3	0.34	0.40	$\boxed{\mathbf{3.06}}$	0.70	0.21
4	0.29	0.37	$\boxed{\mathbf{1.70}}$	0.36	0.11
9	0.59	0.44	0.56	0.79	$\boxed{\mathbf{2.36}}$
12	0.16	$\boxed{\mathbf{2.88}}$	0.97	0.77	0.67
13	0.70	0.09	0.20	0.10	$\boxed{\mathbf{2.82}}$
14	1.17	1.62	0.30	0.26	$\boxed{\mathbf{2.89}}$
15	0.73	0.61	0.57	$\boxed{\mathbf{4.20}}$	0.24
19	0.36	$\boxed{\mathbf{2.23}}$	0.68	0.54	0.40
21	0.80	1.22	$\boxed{\mathbf{2.88}}$	0.07	0.18

3.2 Dataset for TC304 student contest

3.2.1 Description of the data

This subsection applies the proposed method to detect outliers in dataset for TC304 student contest, which is obtained from a clay site. As shown in Table 4, each row represents the data from a certain depth, including liquidity index LI, vertical effective stress σ'$_v$, preconsolidation stress σ'$_p$, molded undrained shear strength S_u^{re}, undrained shear strength S_u. The dataset contains some deliberately modified data. Moreover, geotechnical data itself contain significant uncertainty, which results in great difficulties to the outlier detection.

Table 4. TC304 Student Contest Data

Index	LI	σ'$_{v}$(kPa)	σ'$_{p}$(kPa)	S_u^{re}(kPa)	S_u(kPa)
1	0.98	3.7	13.87	$\boxed{\mathbf{0.88}}$	5.95
2	1.31	7.4	12.95	0.59	4.29
3	1.78	13.87	9.25	0.39	4.07
4	1.51	17.57	17.57	0.39	5
5	1.31	21.27	$\boxed{\mathbf{45.12}}$	0.39	5.95
6	1.34	24.05	21.27	0.59	6.43
7	1.63	27.75	24.05	0.39	7.62
8	1.42	31.45	24.97	0.68	$\boxed{\mathbf{16.7}}$
9	$\boxed{\mathbf{2.52}}$	35.14	29.6	0.68	7.86
10	1.27	39.77	29.6	0.78	12.4
11	1.21	44.39	30.52	0.88	13.1
12	1.38	49.02	36.07	0.98	13.8
13	1.45	$\boxed{\mathbf{51.79}}$	55.49	1.18	17.4
14	1.51	58.27	60.12	$\boxed{\mathbf{1.37}}$	13.1
15	1.22	61.97	48.09	0.98	18.6
16	1.18	66.59	72.14	0.88	17.1
17	0.93	$\boxed{\mathbf{71.21}}$	97.11	1.18	26.2

Table 5 shows the results of the frequency of each suspicious data point in upper 5% MDs in **D**. There are seventeen row vectors occurring in the upper 5% MDs, among which seven rows have the frequency greater than the criterion, i.e. 1/17 = 0.0588. Row vectors of 1, 5, 8, 9, 13, 14, 17 are identified as outliers through RHM-MD with 10^5 times resampling. The number of resampling is much greater than the all probable sample combinations of the 50% data subset to ensure the robustness of the method. The frequency calculated based on the samples from RMH-MD method can indicate the possibility of the outliers. Therefore, the 16-th row vector with 5.51% frequency still have the possibility to be the outlier.

Table 6 shows Euclidean distances of identified outliers, which are calculated based on the procedure proposed in the section 2.2 and used to determine the most probable outlying variable in the outlier row vector. To further illustrate results, two-dimensional scatter plots are drawn in Fig. 4 to show the outliers identified by the proposed approach. As shown in Fig. 4, the identified outliers represented by red cross deviate markedly from other data. It is also shown that the regular points (see circle in Fig. 4) exhibit variability to some extent, but their variability is acceptable.

Table 5. RHM-MD results of the TC304 data

Row index	Counts	Frequency (%)
1	**5790**	**6.81**
2	1100	1.29
3	1282	1.51
4	91	0.11
5	**5544**	**6.52**
6	68	0.08
7	299	0.35
8	**13839**	**16.28**
9	**16293**	**19.17**
10	71	0.08
11	239	0.28
12	1049	1.23
13	**5005**	**5.89**
14	**12408**	**14.60**
15	496	0.58
16	4681	5.51
17	**16745**	**19.70**

4. Conclusion

This paper developed a simple and robust the outlier detection method RHM-MD, which combines the resampling method and Mahalanobis distance. The proposed method can be used to multiple dimensional and small sample problem with considering the correlation

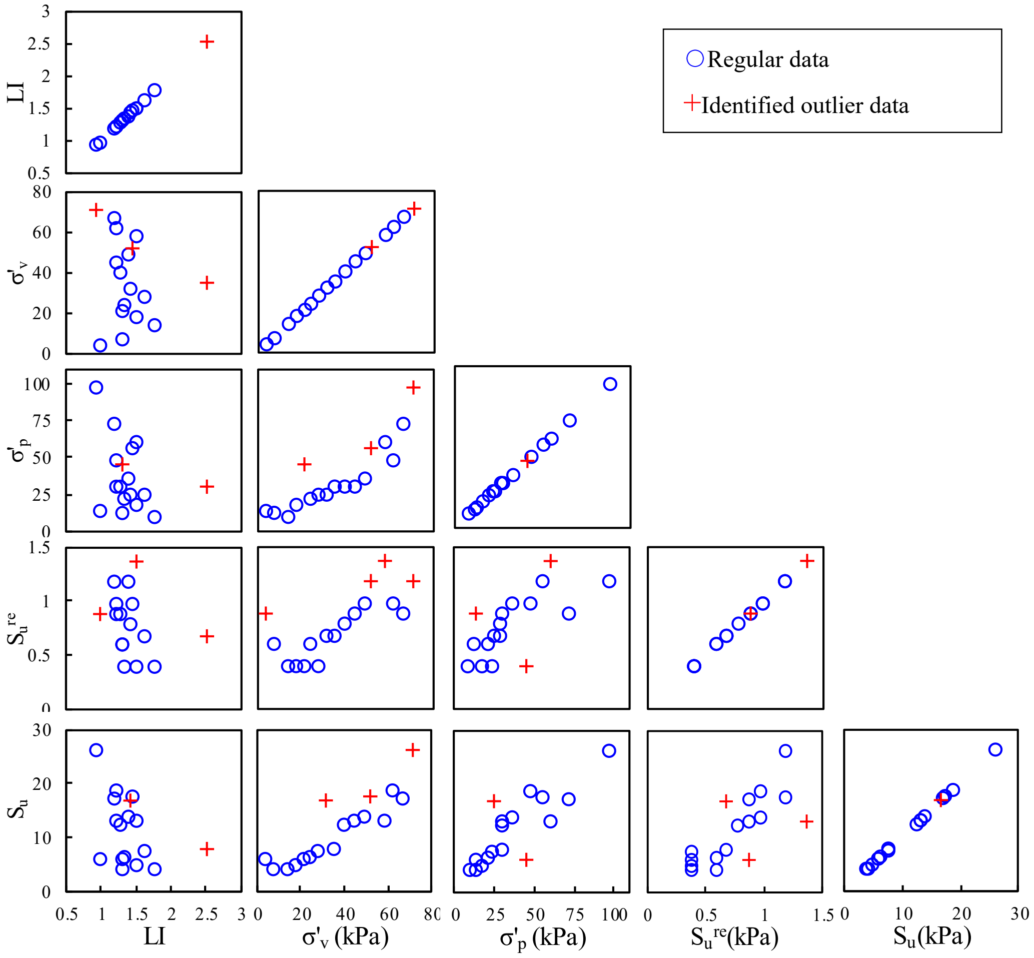

Figure 4. Regular data and identified outliers in the dataset for TC304 student contest

Table 6. Results of TC304 case

Identified outlier index	LI $d^{out}_{\cdot,1}$	σ'_v $d^{out}_{\cdot,2}$	σ'_p $d^{out}_{\cdot,3}$	S_u^{re} $d^{out}_{\cdot,4}$	S_u $d^{out}_{\cdot,5}$
1	25.2	24.6	30.5	**49.5**	29.8
5	7.7	12.9	**40.5**	4.0	6.5
8	0.5	0.3	2.1	2.5	**25.2**
9	**41.4**	3.6	3.2	2.2	10.3
13	8.6	**16.4**	0.6	14.4	1.6
14	12.2	16.6	0.6	**23.1**	21.8
17	4.4	**25.7**	22.5	4.4	4.7

between different variable characteristics. Moreover, the Euclidean distance of each variable is used to determine the most probable outlying variable. A set of five dimensional simulated data, which contains 10 outliers, is used to validate the proposed method. The outlier detection results show that the proposed method is a feasible way to detect the outlying row vector, which provides a robust way for multi-dimensional outlier detection, and can effectively find the most probable outlying variable in the outlier row vector. Finally, the proposed method detected outliers in dataset for TC304 student contest reasonably.

Acknowledgment

The authors thank doctor Zijun Cao (The Wuhan university, Wuhan, China) for his precious comments.

References

Ammann, L.P. 1993. Robust singular value decompositions: a new approach to projection pursuit. *Publications of the American Statistical Association*, 88(422): 505-514.

Barnett, V., and Lewis, T. 1994. *Outliers in Statistical Data 3rd ed*. John Wiley &Sons: New York, 1994.

Egan, W.J., Morgan S.L., and Chem, A. 1998. Outlier detection in multivariate analytical chemical data. *Analytical Chemistry*, 70(11): 2372.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7(1): 1-26.

Frank, E.G. 1969. Procedures for Detecting Outlying Observations in Samples, *Technometrics*, 11(1): 1-21.

Fung, W.K. 1993. Unmasking outliers and leverage points: a confirmation. *Publications of the American Statistical Association*, 88(422), 515-519.

Han, J., and Kamber, M. 2001. *Data Mining*. New York, Morgan Kaufmann Publishers.

Hawkins, D. 1980. *Identification of Outliers*. London, Chapman and Hall.

Hartigan, J.A. 1969. Using Subsample Values as Typical Values. *Publications of the American Statistical Association*, 64(328): 1303-1317.

Li, G.Y., and Chen, Z.L. 1985. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Publications of the American Statistical Association*, 80(391): 759-766.

Mardia, K.V., Kent, J.T. and Bibby, J.M. 1979. *Multivariate Analysis*. Ltd, London.

Mark, H., and Tunnell, D. 1985. Qualitative near-infrared reflectance analysis using Mahalanobis distances. *Analytical Chemistry*, 57: 1449-1456.

Rousseeuw, P.J. 1991. Tutorial to robust statistics. *Journal of Chemometrics*, 5(1): 1-20.

Rousseeuw, P.J., and Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. John Wiley & Sons: New York.

Rousseeuw, P.J., and Zomeren, B.C.V. 1990. Unmasking multivariate outliers and leverage points. *Publications of the American Statistical Association*, 85(411): 633-639.

Walczak, B., and Massart, D.L. 1995. Robust principal components regression as a detection tool for outliers. *Chemometrics & Intelligent Laboratory Systems*, 27(1): 41-54.

William, J.E. and Stephen, L.M. 1998. Outlier Detection in Multivariate Analytical Chemical Data. *Analytical Chemistry*, 70: 2372-2379.

Woodruff, D., and Rocke, D. 1994. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Publications of the American Statistical Association*, 89(427), 888-896.

Yuen, K.V., and Mu, H.Q. 2012. A novel probabilistic method for robust parametric identification and outlier detection. *Probabilistic Engineering Mechanics,*30(4), 48-59.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*. 2(1): 49–55.