Hydraulic and Civil Engineering Technology VII
M. Yang et al. (Eds.)

doi:10.3233/ATDE220867

The Influence Study of Statistical Uncertainty on Probabilistic Outlier Detection for Geotechnical Engineering Data

Shuo ZHENG [a], Jinqiu LIANG [a], Xiaoquan GAN [a], Huifeng ZHENG [a,1] and Ning LIU [a]
[a] *Huadong Engineering Corporation Limited of Power China Hangzhou, China*

Abstract. Outlier is attached importance in statistics and engineering, because it might result in misleading identification results. However, there is significant uncertainty in the outlier detection, when an outlying observation lies close to the boundary between outliers and regular data or there are sparse observations. The associated uncertainty of outlier mostly results from statistical uncertainty of parameters, such as mean value and standard deviation. However, it is unknown how the statistical uncertainty influences the outlier detection. This paper compares two outlier detection methods for influence study of statistical uncertainty on probabilistic outlier detection. One is based on Mahalanobis distance (MD) using the total probability theorem combining with the half-means method (RHM). The other is RHM method with Bayesian machine learning (BML), which can consider the statistical uncertainties of parameters in MD. The simulated dataset with outliers are used to comparative study. Different dimensional dataset and various numbers of observations and outliers are simulated. Thereinto, outliers are simulated through double-mode triangle distribution. The results show that it is necessary to consider the statistical uncertainty for sparse multivariate observations.

Keywords. Outlier detection, geotechnical engineering, resampling by half-means, statistical uncertainty

1. Introduction

An outlier is one that appears to deviate markedly from other observations [1]. 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 [2-3]. In practice, directly incorporating measurements or observations with outliers into data analysis might lead to significant bias. It is necessary to detect outliers that mix in the regular data patterns [4], so as to eliminate them and reduce their impacts on statistical inferences [5].

There are many outlier detection methods developed based on different models such as statistical models [6-7], regression models [8-11] and classification models [12-

[1] Huifeng Zheng, Corresponding author, Huadong Engineering Corporation Limited of Power China Hangzhou, China; E-mail: zheng_hf@hdec.com.

13] for different problems of various regions. Among of statistical models, multivariate normal model is usually used to characterize the geotechnical parameters (e.g., [14-15, 16]), which are often transformed into standard normal variables to analysis (e.g., [17-18]), because it is mathematically tractable. The traditional Gaussian model based statistical outlier detection techniques have been developed based on the distance of a data instance to the estimated mean. A threshold is applied to the anomaly scores to determine the outliers. The Mahalanobis distance incorporates the dependencies between the variables. A cutoff value of the Mahalanobis distance for outlier identification can be chosen for the performer [8,19]. However, these previous methods ignored the associated uncertainties in outlier detection.

If the data contains huge variability that tends to be like the actual outliers, it will be difficult to distinguish [20]. It is also common that data analysis (e.g., geotechnical design) is conducted under limited geotechnical site investigation data (e.g., such as the cone penetration test measurements), which maybe contain outliers. Especially, measurements or observations are often sparsely available in geotechnical investigation, so there is a significant uncertainty in statistical estimations and outlier detection. Therefore, a rigorous and robust probabilistic approach for outlier detection, which can consider the associated uncertainties rationally and find outlying components in the outlying row vector, was proposed by [21].

In this paper, two outlier detection methods are compared for influence study of statistical uncertainty on probabilistic outlier detection. One is based on Mahalanobis distance (MD) using the total probability theorem combining with the resampling by half-means method (RHM). The other is RHM combining with Bayesian machine learning (BML), which can consider the statistical uncertainties of parameters in MD. The simulated datasets with outliers are used to comparative study. Different dimensional datasets and various numbers of outliers are simulated. The outliers and regular data are simulated through double-mode triangle distribution and multivariate Gaussian distribution, respectively.

2. Heading Outlier Detection

The probabilistic outlier detection method in this paper can be conducted through two parts. Firstly, the outlying row vectors are identified based on the Mahalanobis distance (MD) [22, 19] and the uncertainty of them is quantified using the RHM combining with total probability theorem. The row vector, of which outlying probability is greater than 0.5, can be considered a possible outlying row vector. Secondly, outlying components in the possible outlying row vectors are detected by an exclusion method. The details of above two methods are introduced in following section.

Let $\mathbf{X}_{\mathrm{E}} \in \mathbb{R}^{N\times d}$ denote the entire dataset, which are geotechnical in-situ test measurements, and it contains N row vectors $\mathbf{X}_i$, i =1, 2, …, N. Each row vector $\mathbf{X}_i$ contains d components $x_{i,j}$, j = 1, 2, …, d. Calculation details of $P(\mathbf{X}_i \in \Omega^d_{out})$ (i.e., the outlying probability of $\mathbf{X}_i$) are provided in the next section. Herein, the threshold probability of $P(\mathbf{X}_i \in \Omega^d_{out})$ is taken as 0.5, which was also adopted in the literature (e.g., [10, 11]). The exclusive method was proposed by [21] to detect the outlying components. The differential of the $P(\mathbf{X}_{i,\bullet} \in \Omega^d_{out})$ and $P(\mathbf{X}_{i,\bullet\setminus j} \in \Omega^{d-1}_{out})$ (i.e., $\Delta P_{i,j}$ = $P(\mathbf{X}_i \in \Omega^d_{out})$ - $P(\mathbf{X}_{i,\setminus j} \in \Omega^{d-1}_{out})$; $\mathbf{X}_{i,\setminus j} = [x_{i,1}, x_{i,2}, ..., x_{i,j-1}, x_{i,j+1}, ..., x_{i,d}]$, i = 1, 2, …, N) is used to represent the contribution of the $x_{i,j}$ to the outlying row vector, $\mathbf{X}_{i,\bullet} \in \Omega^d_{out}$ .The

maximum of $\Delta P_{i,j}$ is considered as the most probable outlier component $x_{i,j}$ in the ith outlying row vector $\mathbf{X}_{i,\bullet}$. The exclusive procedure is repeatedly performed for each component and each row vector.

3. Methodology

In this paper, two probabilistic outlier detection methods are compared for the influence study of statistical uncertainty. "Method-1" is based on Mahalanobis distance (MD) using the total probability theorem combining with the half-means method (RHM). "Method-2" is Method-1 with Bayesian machine learning (BML), which can consider the statistical uncertainties of parameters in MD [21].

3.1. Method-1

William and Stephen 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 [23]. The theoretical connotation is that half of the data must be regular, otherwise it is meaningless to search for outliers. The thought of the method is adopted in this paper. Firstly, the possible 50% regular dataset is detected by examining the MDs obtained from sampling without replacement from the original data set. Then repeat above sampling with replacement. However, outlying probability was not quantified. The probability of outlier can be quantified by following Method-1.

Initially, select the subset of observations without replacement from the input entire sample matrix $\mathbf{X}_{\mathrm{E}}$ until up to the size of a half of the $\mathbf{X}_{\mathrm{E}}$, which is denoted as $\mathbf{X}_{sample,\,k}$ (i.e., k = 1, ..., N_{re}, where N_{re} is the number of samplings with replacement), and the above resampling is conducted N_{re} times. Based on the total probability theorem, the probability of outlier for each row vector can be estimated efficiently using conditional probabilities of outlier.

$$P\left(\mathbf{X}_{i,\mathrm{g}} \in \Omega_{out}^{d}\right) = \sum_{k=1}^{Nre} I\left(\mathbf{X}_{i,\mathrm{g}} \in \Omega_{out}^{d} \mid \mathbf{X}_{sample,k}\right) P\left(\mathbf{X}_{sample,k}\right) \quad for\left(i = 1,2,...,N\right) \tag{1}$$

where $\mathbf{X}_{i,\bullet} \in \mathbb{R}^{1\times d}$ is the ith row vector; Ω^{d}_{out} is the subset of outliers in the d-dimensional space; N_{re} is the number of resampling; $I(\mathbf{X}_{i,\bullet} \in \Omega^{d}_{out}|\mathbf{X}_{sample,\,k})$ is an indicator function. If $MD_{i,\bullet}$ is greater than $\sqrt{\Phi^{-1}_{\chi^2_d}(0.975)}$, $I(\mathbf{X}_{i,\bullet} \in \Omega^{d}_{out}|\mathbf{X}_{sample,\,k})$ is equal to one; otherwise, it is taken as zero. $MD_{i,\bullet}$ is the Mahalanobis distance of $\mathbf{X}_{i,\bullet}$ based on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Mahalanobis 1936). And the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the $\mathbf{X}_{sample,\,k}$. $P(\mathbf{X}_{sample,\,k})$ is the probability of $\mathbf{X}_{sample,\,k}$, which is equal to 1/ N_{re}, because the N_{re} possible samples $\mathbf{X}_{sample,\,k}$ (i.e., k = 1, ..., N_{re}) are considered having the same probability during the resampling process. If $P(\mathbf{X}_{i,\bullet} \in \Omega^{d}_{out})$ is greater than 0.5, $\mathbf{X}_{i,\bullet}$ will be considered a possible outlying row vector denoted as $\mathbf{X}^{out}{}_{i,\bullet}$ [10].

3.2. Method-2

It is hence important to estimate **μ** and **Σ** based on the regular data set to detect outliers, which are unknown before analysis. To consider the uncertainty of the parameters (e.g., mean value μ and covariance matrix Σ) in the Mahalanobis distance, Bayesian machine learning (BML) is used to quantify posterior uncertainty of multivariate normal model parameters by generating the posterior samples of the parameters. Based on the total probability theorem, the probability of outlier for each row vector can be estimated efficiently using conditional probabilities of outlier:

$$P\left(\mathbf{X}_{i,\mathrm{g}} \in \Omega_{out}^{d}\right) = \sum_{k=1}^{Nre} P\left(\mathbf{X}_{i,\mathrm{g}} \in \Omega_{out}^{d} \mid \mathbf{X}_{sample,k}\right) P\left(\mathbf{X}_{sample,k}\right) \quad for\left(i = 1,2,...,N\right) \tag{2}$$

where $P(\mathbf{X}_{i,\bullet} \in \Omega^{d}_{out} | \mathbf{X}_{sample,\ k})$ is the conditional probability of outlier for each row vector in the **X** for given $\mathbf{X}_{sample,\ k}$. Similarly, the conditional probability of outlier for each row vector $P(\mathbf{X}_{i,\bullet} \in \Omega^{d}_{out} \mid \mathbf{X}_{sample,\ k})$ can be estimated efficiently using the conditional probability density function (PDF) of **μ** and **Σ,** i.e., $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}_{sample,\ k})$, based on the total probability theorem:

$$P\left(\mathbf{X}_{i,\mathrm{g}} \in \Omega_{out}^{d} \mid \mathbf{X}_{sample,k}\right) = \int I\left(MD_{i,\mathrm{g}} > \sqrt{\Phi^{-1}_{\chi^2_d}\left(0.975\right)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{X}_{sample,k}\right) d\boldsymbol{\mu} d\boldsymbol{\Sigma} \tag{3}$$

where $I\left(MD_{i,\mathrm{g}} > \sqrt{\Phi^{-1}_{\chi^2_d}\left(0.975\right)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is an indicator function. If $MD_{i,\bullet}$ is greater than $\sqrt{\Phi^{-1}_{\chi^2_d}\left(0.975\right)}$, $I\left(MD_{i,\mathrm{g}} > \sqrt{\Phi^{-1}_{\chi^2_d}\left(0.975\right)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is equal to one; otherwise, it is taken as zero. In other word, if $MD_{i,\mathrm{g}} > \sqrt{\Phi^{-1}_{\chi^2_d}\left(0.975\right)}$, $\mathbf{X}_{i,\bullet}$ can be considered as an outlier for a given **μ** and **Σ**. The posterior distribution of **μ** and **Σ**, $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}_{sample,\ k})$, which can be derived based on Bayesian framework. The posterior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}_{sample,\ k})$ contains the statistical uncertainty of **μ** and **Σ**. However, for lack of space, the Bayesian framework for parameters identification is not introduced with details. Readers can refer to [21].

4. Influence Study Using Simulated Dataset with Outliers

4.1. Simulating Data

To explore the effect of the number of observations, outliers, and dimensions (i.e., the number of attributes) on the proposed method, 25 cases of simulated data with various number of observations **X** and outliers for different dimensional spaces (e.g., 3, 5, 7 and 9 dimensions) are drawn from the multi-dimensional normal distribution and the triangular distribution, simultaneously, each of them concludes 30 runs; simulation for each run is independent and random as shown in figure 1. The number of observations is set as 20, 40, 60, 80 and 100. Given the number of observations, the number of outliers is set as 2, 4, 6, 8 and 10.

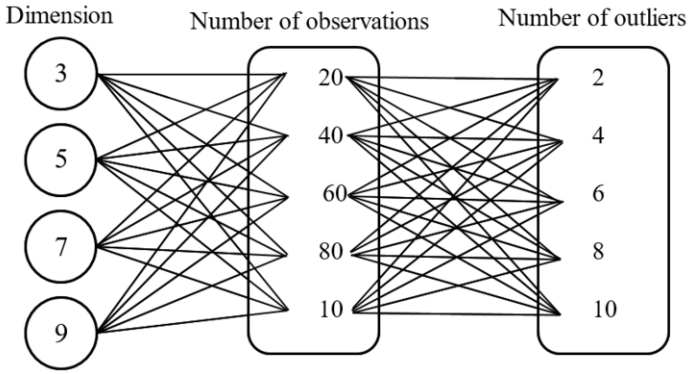


Figure 1. Summary of simulated dataset cases.

4.2. Comparative Study

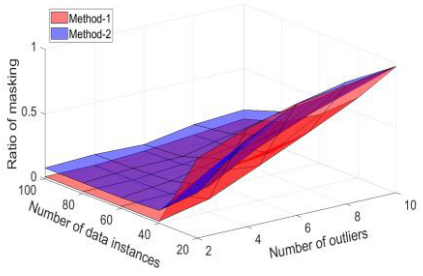
The Method-1 and Method-2 are applied to simulated data with N_{re} = 500. Two indicators are used to quantify the performance of outlier detection:

$$r_{masking} = \frac{N_{masking}}{N_{out}} \tag{4}$$

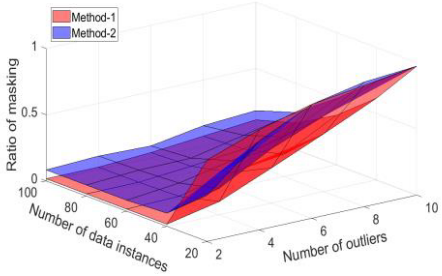
where $r_{masking}$ is the ratio of masking; $N_{masking}$ denotes the number of real outlying row vectors not being detected; N_{out} is the number of real outlying row vectors.

$$r_{swamping} = \frac{N_{swamping}}{N_{regular}} \tag{5}$$

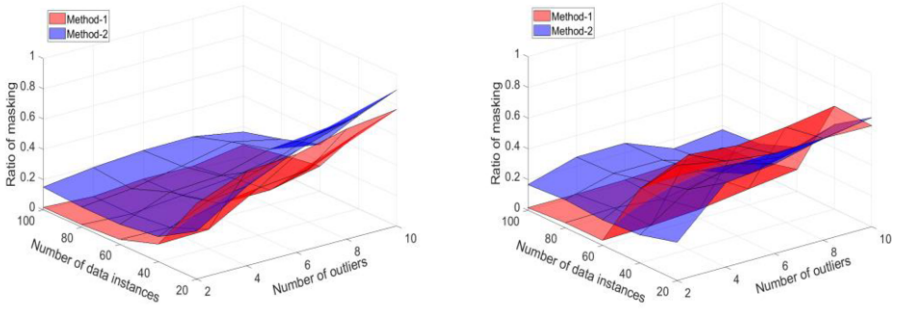
where $r_{swamping}$ is the ratio of swamping; $N_{swamping}$ denotes the number of real regular row vectors mistakenly identified as outlying row vectors; $N_{regular}$ is the number of real regular row vectors. It is desirable to obtain low values of both indicators [10].



(a) 3 Dimension



(b) 5 Dimension

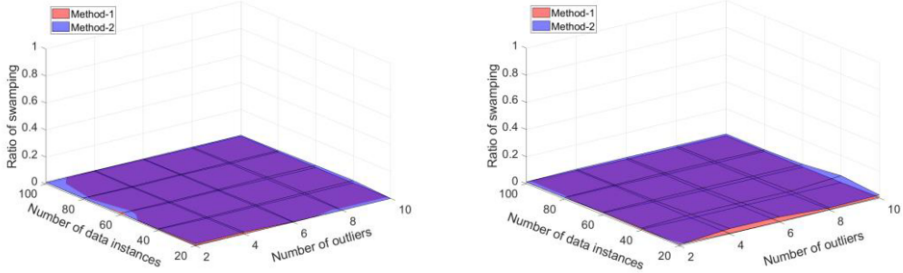


(c) 7 Dimension (d) 9 Dimension

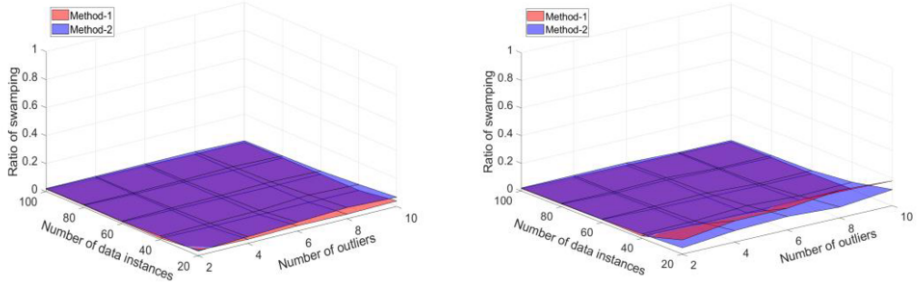
Figure 2. Results of the average masking ratio using two methods.

Each case considered in this section uses 30 simulated datasets. The averaged r_{masking} of 30 runs for each case in four different dimensions based on the proposed approach are summarized in figure 2. As shown in figure 2, the results of Method-1 and Method-2 are in red and blue, respectively. The r_{masking} results show that r_{masking} of Method-1 is greater than r_{masking} of Method-2, when the number of observations is 20 and the number of outliers is less than 6. Especially, for the 9-dimension observations, r_{masking} of Method-1 is greater than r_{masking} of Method-2, when the number of observations is less than 60, as shown in figure 2(d). The poor performance of the Method-1 for the small number of data cases indicates that the μ and Σ estimators based on the less observations exist great statistical uncertainty. Therefore, the statistical uncertainty should be considered for outlier detection for sparse multivariate observations. When the number of observations is enough to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ exactly, simpler Method-1 is recommended because of its efficiency.

Although the $r_{masking}$ is higher for the case of high ratio of outlier to observation, the $r_{swamping}$ of them is lower than 0.1 for almost cases by two methods as shown in figure 3. As shown in figure 3, Method-1 and Method-2 both can perform well (i.e., low values of both indicators) for most cases of multi-dimensional observation.



(a) Dimension (b) 5 Dimension



(c) 7 Dimension (d) 9 Dimension

Figure 3. Results of the average swamping ratio using two methods.

5. Conclusion

This paper compared two probabilistic outlier detection methods for the influence study of statistical uncertainty. The Method-2 was proposed by [21]. The Method-1 is based on RHM without considering statistical uncertainty of parameters. Method-1 is simpler than Method-2. So, the computation time under the same circumstance shows that the Method-1 significantly shortens the computational cost. The above methods were applied to the same simulated dataset. Comparison of outlier detection results through two indicators (i.e., the ratio of masking and the ratio of swamping) indicates that the statistical uncertainty is necessarily considered in outlier detection for sparse multivariate observations.

References

[1] Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics. 1969;11(1):1-21.
[2] Hawkins D. Identification of outliers. London: Chapman and Hall.1980.
[3] Barnett V, Lewis T. Outliers in Statistical Data 3rd ed. John Wiley &Sons: New York. 1994.
[4] Han J, Kamber M. Data Mining. New York, Morgan Kaufmann Publishers. 2001.
[5] Rousseeuw PJ. Tutorial to robust statistics. Journal of Chemometrics. 1991;5(1):1-20.
[6] Hodge V, Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review. 2003;22: 85-126.
[7] Markou M, Singh S. Novelty detection: a review: Part 1: statistical approaches. Signal Processing. 2003; 83(12):2481-2497.
[8] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. John Wiley & Sons: New York. 1987.
[9] Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011; 1(1):73-79.
[10] Yuen KV, Mu HQ. A novel probabilistic method for robust parametric identification and outlier detection. Probabilistic Engineering Mechanics.2012;30(4):48-59.
[11] Yuen KV, Ortiz GA. Outlier detection and robust regression for correlated data. Computer Methods in Applied Mechanics and Engineering. 2017;313:632-646.
[12] Japkowicz N, Myers C, Gluck MA. A Novelty Detection Approach to Classification. Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95). 1995; p. 518–523.
[13] Zhang Y, Meratnia N, Havinga PJ. Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials'2010; 12(2):159-170.
[14] Ching J, Phoon KK. Modeling parameters of structured clays as a multivariate normal distribution. Canadian Geotechnical Journal. 2012;49(5):522-545.

[15] Ching J, Phoon KK and Chen CH. Modeling CPTU parameters of clays as a multivariate normal distribution. Canadian Geotechnical Journal. 2104;51(1):77-91.
[16] Ching J, Phoon KK, Li DQ. Robust estimation of correlation coefficients in the multivariate normal framework. Structural Safety. 2016;63:21-32.
[17] Liu PL, Der Kiureghian A. Multivariate distribution models with prescribed marginals and covariances. Probabilistic Engineering Mechanics. 1986; 1(2):105-112.
[18] Li DQ, Wu SB, Zhou CB, Phoon KK. Performance of translation approach for modeling correlated non-normal variables. Structural Safety. 2012;39:52-61.
[19] Rousseeuw PJ, Zomeren BCV. Unmasking multivariate outliers and leverage points. Publications of the American Statistical Association.1990;85(411):633-639.
[20] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys (CSUR). 2009;41(3):15.
[21] Zheng S, Zhu YX, Li DQ, Cao ZJ, Deng QX& Phoon KK. Probabilistic outlier detection for sparse multivariate geotechnical site investigation data using Bayesian learning. Geoscience Frontiers. 2021; 12(1):425-439.
[22] Mahalanobis PC. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India. 1936; 2(1): 49–55.
[23] William JE, Stephen LM. Outlier detection in multivariate analytical chemical data. Analytical Chemistry. 1998;70: 2372-2379.