2018 TC304 Student Contest
17-19 Aug 2018, Harbin, China

Outlier Detection for Multivariate and Sparse Site-specific Geotechnical Data

Te Xiao[1] and Jian He[2]

[1]*Wuhan University. Email: short_xiaote@whu.edu.cn*
[2]*Wuhan University. Email: hejian@whu.edu.cn*

Abstract: Site characterization is usually carried out based on geotechnical data from site investigation. However, the existence of outliers in geotechnical data might lead to an incorrect characterization result, which necessitate the outlier detection. Site-specific geotechnical data is usually multivariate, sparse and might has a certain trend. This study proposed an outlier detection algorithm that considers the influence of statistical uncertainty caused by limited data through Bayesian method. Posterior samples of model parameters are generated by a Bayesian updating method together with subset simulation. Based on the posterior samples, the probability of outlier can be obtained, both block-wise and component-wise. The proposed algorithm is applied to a dataset from a clay site with some artificial outliers. The results show that the outliers can be effectively identified by the proposed outlier detection algorithm.

Keywords: Outlier detection; Bayesian method; BUS; Mahalanobis distance; Abnormity degree.

1. Introduction

Anomalies in geotechnical data are inevitable and have great impacts on site characterization. Identification of these outliers will help to improve the quality of data. Various algorithms have been proposed to identify the outliers from observations, such as 3-σ rule, Mahalanobis distance (Rousseuw and Leroy, 1987), local outlier factor (Breunig et al., 2000), and Bayesian method (Yuen and Mu, 2012).

In geotechnical practice, the site-specific geotechnical data is usually sparse and multivariate. On the one hand, most of outlier detection algorithms that are developed for large dataset, such as the machine learning technique-based algorithms, do not work well for the sparse geotechnical data. Furthermore, the sparse site-specific geotechnical data is always associated with significant statistical uncertainty. It is necessary to incorporate the statistical uncertainty into outlier detection.

On the other hand, different geotechnical parameters can be obtained at the same location, and they are correlated to a certain degree (Ching and Phoon, 2012). Existing outlier detection algorithms for multivariate data mainly focus on the block-wise outlier detection, which aims to identify abnormal multivariate data points as a whole. However, an abnormal multivariate data point does not mean all the components of the data point are abnormal. To fully utilize the limited data for site characterization, it is more desirable to identify the value of which parameter within the outlier point is abnormal, namely the component-wise outlier detection.

This study proposes an outlier detection algorithm for multivariate and sparse site-specific geotechnical data, in which the influence of statistical uncertainty would be properly considered through Bayesian method. The proposed algorithm not only gives the possibility for each identified outlier, but also suggests the value that is the most possible anomaly within an outlier point.

2. Bayesian inference

2.1 Multivariate model

Geotechnical parameters are typically non-normal because most of them are physically nonnegative. For simplicity, it is assumed in this study that the geotechnical parameters are lognormally distributed (Ching and Phoon, 2012). In other words, the logarithms of geotechnical parameters are assumed to have a multivariate normal distribution.

Due to the fact that some geotechnical parameters are related to the depth or the effective vertical stress, σ'_v, such as the undrained shear strength, s_u, a linear regression model between σ'_v and other geotechnical parameters can be used to detrend the data. Consider that D geotechnical parameters (σ'_v is excluded) for example. The multivariate model is expressed by

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\boldsymbol{y} = [y_1, y_2, \ldots, y_D]$ is the logarithms of geotechnical parameters; $\boldsymbol{x} = [1, z]$ and z is the logarithm of σ'_v; $\boldsymbol{\beta}$ is a 2-by-D matrix of regression parameters; $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_D]$ is the regression residual that follows a zero-mean multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Eventually, the multivariate geotechnical data is described by parameters ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$), and the two model parameters can be characterized through Bayesian method in the following subsections.

2.2 Posterior of model parameters

For sparse site-specific data, the influence of statistical uncertainty on model parameters ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$) should be considered to robustly identify outliers. This can be achieved through the Bayesian method.

According to the multivariate model expressed as Eq. (1), the posterior probability of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be obtained by the Bayes' rule:

$$P(\boldsymbol{\beta},\Sigma \mid \boldsymbol{X},\boldsymbol{Y}) = k^{-1}P(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{\beta},\Sigma)P(\boldsymbol{\beta})P(\Sigma) \qquad (2)$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^{\mathrm{T}}$ is N observations of $\boldsymbol{x}$ and $\boldsymbol{x}_n = [1, z_n]$ ($n = 1, 2, \ldots, N$); $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N]^{\mathrm{T}}$ is N observations of $\boldsymbol{y}$ and $\boldsymbol{y}_n = [y_{n1}, y_{n2}, \ldots, y_{nD}]$; k is a normalizing constant independent from model parameters; $P(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma})$ is the likelihood function; $P(\boldsymbol{\beta})$ and $P(\boldsymbol{\Sigma})$ are prior probabilities of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, respectively. Herein, the model parameters are assumed to be independent.

For the multivariate model defined by Eq. (1), the likelihood function can be calculated as

$$P(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{\beta},\boldsymbol{\Sigma}) = \prod_{n=1}^{N}(2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{x}_n\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{x}_n\boldsymbol{\beta})^{\mathrm{T}}\right] \qquad (3)$$

Its logarithm equals

$$\begin{aligned}\ln P(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{\beta},\boldsymbol{\Sigma}) = &-\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| \\ &-\frac{1}{2}\mathrm{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\right]\end{aligned} \qquad (4)$$

With respect to the prior, considering the case with little prior information, diffused distributions, e.g., uniform distributions with relatively wide ranges, can be selected as prior distribution to represent the non-informative prior information. For convenience, the lower triangle matrix $\boldsymbol{L}$ from the Cholesky decomposition of $\boldsymbol{\Sigma}$ (i.e., $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$) is used instead to parameterize the multivariate correlation among geotechnical parameters. All elements in $\boldsymbol{\beta}$ and $\boldsymbol{L}$ are assumed to be independent and identically distributed.

When there is no closed form for the posterior probability, the posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{L}$ can be numerically generated by Markov chain Monte Carlo simulation (MCMCS). However, since the model parameters in this study is high-dimensional, the commonly-used MCMCS algorithms (e.g., Metropolis-Hastings algorithm) become inefficient and difficult to achieve the stationary state. To efficiently draw posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{L}$, an algorithm named Bayesian updating with structural reliability methods (BUS) (Straub and Papaioannou, 2014) is employed in this study.

2.3 Generating posterior samples using BUS

The core idea of BUS is to convert a Bayesian updating problem to an equivalent reliability analysis problem. In reliability analysis problems, the occurrence probability of failure event, i.e., the failure probability, is of interest. In the context of BUS, the performance function in the equivalent reliability analysis problem is defined as (Straub and Papaioannou, 2014)

$$F = \ln U - \ln c - \ln P(\boldsymbol{Y},\boldsymbol{X} \mid \boldsymbol{\beta},\boldsymbol{\Sigma}) \qquad (5)$$

where U is a random variable that uniformly distributed on [0, 1]; c is a positive scaling constant that ensures $cP(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma}) \leq 1$. To satisfy the constraint, c is taken as the reciprocal of the maximum likelihood in this study.

In the equivalent reliability analysis problem, the failure event is defined as $\{F < 0\}$ with $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ distributed as $P(\boldsymbol{\beta})$ and $P(\boldsymbol{\Sigma})$, respectively, and

U uniformly distributed on [0, 1]. It can be proven that the failure samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ conditional on $\{F < 0\}$ follow the posterior distribution (Straub and Papaioannou, 2014).

To solve the equivalent reliability analysis problem and obtain the failure samples, simulation-based reliability method, such as direct Monte Carlo simulation and subset simulation (Au and Beck, 2001), can be used. Considering the dimensionality of parameters and the sampling efficiency, the subset simulation is adopted in this study. Interested readers can be referred to the literature (Au and Beck, 2001; Straub and Papaioannou, 2014) for detailed algorithm description and implementation procedure of BUS with subset simulation.

2.4 Maximum likelihood estimator

Since scaling constant c in the BUS depends on the maximum likelihood, the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are required. They can be obtained easily by taking the differentiation of log-likelihood function (i.e., Eq. (4)) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, namely

$$\frac{\partial \ln P(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{\beta},\Sigma)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X\beta})\Sigma^{-1} = \mathbf{0} \tag{6}$$

$$\begin{aligned}\frac{\partial \ln P(\boldsymbol{X},\boldsymbol{Y} \mid \boldsymbol{\beta},\Sigma)}{\partial \Sigma} &= -\frac{N}{2}\Sigma^{-1} \\ &+\frac{1}{2}\Sigma^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X\beta})\Sigma^{-1} = \mathbf{0}\end{aligned} \tag{7}$$

By this means, the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are evaluated as

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\right) \tag{8}$$

$$\hat{\Sigma} = \frac{1}{N}(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X\beta}) \tag{9}$$

With the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, the maximum likelihood can be directly calculated with ease.

3. Outlier detection

3.1 Block-wise outlier detection

Through the Bayesian inference, a large number of posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are obtained. For a given sample set of ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$), the Mahalanobis distance can be used as a measure to perform block-wise outlier detection.

The Mahalanobis distance, d, is defined as (Rousseuw and Leroy, 1987)

$$d = \sqrt{(\boldsymbol{y}-\boldsymbol{x\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{x\beta})^{\mathrm{T}}} \tag{10}$$

For a multivariate normal distribution, the probability density of an observation is uniquely determined by the Mahalanobis distance. Although d^2 is chi-squared distributed with D degrees of freedom in general, the criterion is improper when the amount of data is limited.

Instead, this study finds the most possible outlier that has the largest Mahalanobis distance, for a given sample of ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$). Considering the impact of statistical uncertainty, the possibility of one point being the most possible outlier can be evaluated by traversing all posterior samples of ($\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$). Eventually, the points with outlier probability greater than a prescribed threshold (e.g., 0.01 in this study) are recognized as the block-wise outliers.

3.2 Component-wise outlier detection

Most of existing outlier detection algorithms only identify the abnormal data points. With these methods, the identified outlier points are discarded even though there might be only one variable's value abnormal within the data points. Considering the cost of site-investigation, those normal data should be fully utilized for site characterization. The target of component-wise outlier detection is to identify which geotechnical parameter has an abnormal value that makes the point an outlier. If one abnormal value is identified, then this value can be replaced by the expectation conditional on the measurements of other geotechnical parameters within the data point.

Generally, abnormal values are less probable than normal values. Therefore, it is assumed that there is only one variable with an abnormal value for an identified outlier point. The abnormity degree (AD) of the i-th variable's value within the n-th data point is defined as

$$AD = 1 - \frac{P(\boldsymbol{y}_n,\boldsymbol{x}_n \mid \boldsymbol{\beta},\Sigma)}{P(\boldsymbol{y}_n^{(i)},\boldsymbol{x}_n \mid \boldsymbol{\beta},\Sigma)} \tag{11}$$

where $\boldsymbol{y}_n^{(i)} = [y_{n1},\ldots,\bar{y}_{ni},\ldots,y_{nD}]$; $\bar{y}_{ni}$ is the expectation conditional on the measurements of other geotechnical parameters in $\boldsymbol{y}_n$, i.e., $\bar{y}_{ni}$ = $\mathrm{E}\left(y_{ni} \mid y_{n1},\ldots,y_{n,i-1,},y_{n,i+1},\ldots,y_{n\mathrm{D}},\boldsymbol{x}_n,\boldsymbol{\beta},\Sigma\right)$. Herein, it is assumed that $\boldsymbol{x}_n$ is always normal. If y_{ni} is an anomaly, and the measurements of other geotechnical parameters are inliers, then $\boldsymbol{y}_n^{(i)}$ can be regarded as the corrected data. The more abnormal the *i*-th variable's value is, the smaller the probability of original data is compared to the probability of the corrected data, hence a larger *AD*. Therefore, the variable with the largest *AD* can be identified as the abnormal value within the data point.

For each outlier identified by block-wise outlier detection, *AD* of each variable is calculated given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, and the variable with the largest *AD* is denoted as the anomaly given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Similarly, repeatedly performing the component-wise outlier detection for all posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, and then the probability of the component-wise outlier can be obtained.

4. Implementation procedure

The implementation procedure of the proposed outlier detection is as follows:

(1) Take logarithm of the original geotechnical data, and construct $\boldsymbol{X}$ with $\ln(\sigma'_v)$ and $\boldsymbol{Y}$ with the rest geotechnical parameters.

(2) Find the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ using Eqs. (8)-(9), and calculate the maximum likelihood.

(3) Generate posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ using BUS with subset simulation as described in Section 2.3.

(4) Repeat the block-wise outlier detection (see Section 3.1) for each posterior sample of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and calculate the probability of the outlier point. The points with outlier probability greater than 0.01 are recognized as the block-wise outliers.

(5) Repeat the component-wise outlier detection (see Section 3.2) for each posterior sample of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ and calculate the probability of the abnormal variable. The variable with the largest outlier probability within the abnormal date point identified in Step (4) is considered as the component-wise outliers.

5. Illustrative example

The proposed outlier detection for multivariate and sparse site-specific geotechnical data is applied to a geotechnical dataset from a clay site in the ISSMGE TC304 database (Ching and Phoon, 2012), as shown in Table 1. Each data point contains five geotechnical parameters measured at a certain depth, including liquidity index (LI), vertical effective stress (σ'_v), preconsolidation stress (σ'_p), remolded undrained shear strength (s_u^{re}), and undrained shear strength (s_u). Four values are replaced by outliers as marked with a star in Table 1.

Table 1. Geotechnical dataset from a clay site.

ID	LI	σ'_v (kPa)	σ'_p (kPa)	s_u^{re} (kPa)	s_u (kPa)
1	0.98	3.70	13.87	0.88	5.95
2	1.31	7.40	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	5.00
5	1.31	21.27	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	16.74*
9	2.52*	35.14	29.60	0.68	7.86
10	1.27	39.77	29.60	0.78	12.38
11	1.21	44.39	30.52	0.88	13.10
12	1.38	49.02	36.07	0.98	13.81
13	1.45	51.79	55.49	1.18	17.38
14	1.51	58.27	60.12	1.37	13.10
15	1.22	61.97	48.09	0.98	18.57
16	1.18	66.59	72.14	0.88	17.14
17	0.93	71.21	97.11	1.18	26.19

Note: LI = liquidity index; σ'_v = vertical effective stress; σ'_p = preconsolidation strength; s_u^{re} = remolded undrained shear strength; s_u = undrained shear stress; * represents the artificial outliers.

During the Bayesian inference, the scaling factor *c* in BUS can be readily obtained using the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ given by Eqs. (8)-(9). For this example, $\ln c = -0.3644$. Subset simulation is then applied to generate posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, with the conditional probability set as 0.1, and the number of samples in each level set as 50,000. Eventually, 461,070 posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are obtained.

Applying the proposed outlier detection to the data in Table 1 based on the generated posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, the block-wise outliers can be identified as shown in Table 2. With a threshold of 0.01, the outlier points identified are point no. 5, 9, 14 and 8, in a descending probability order. Among all artificial outliers (point no. 3, 5, 8 and 9), point no. 5, 8 and 9 are correctly identified. Point no. 3 is not declared as an outlier, which is because the outlier only slightly deviates from the norm.

Table 2. Probability of block-wise outlier.

ID	Probability	ID	Probability
1	0.0000	10	0.0000
2	0.0000	11	0.0000
3	0.0059	12	0.0000
4	0.0000	13	0.0000
5	**0.4684**	14	**0.0495**
6	0.0000	15	0.0000
7	0.0000	16	0.0000
8	**0.0176**	17	0.0026
9	**0.4559**		

Then, implement the component-wise outlier detection algorithm for each identified outlier point. The probability of component-wise outlier is listed in Table 3. Outliers in data no. 5, 8 and 9 are successfully identified. Although there is no artificial outlier in data no. 14, it is still identified as an outlier. Note that the identified component-wise outlier of no. 14 is the largest value among the observations of s_u^{re}. Therefore, it is reasonable to consider it as an outlier to some extent.

Table 3. Probability of component-wise outlier.

ID	LI	σ'_P	s_u^{re}	s_u
5	0.0000	**0.6128**	0.3871	0.0000
8	0.0000	0.0003	0.0000	**0.9997**
9	**1.0000**	0.0000	0.0000	0.0000
14	0.0003	0.0117	**0.9851**	0.0002

6. Conclusions

Anomalies in geotechnical data are inevitable and have great impacts on site characterization. Site-specific geotechnical data is usually multivariate, sparse and might has a certain trend. This study proposed an outlier detection algorithm that considers the influence of statistical uncertainty caused by limited data through Bayesian method. A multivariate normal model with linear regression is used to describe the geotechnical data. Posterior samples of model parameters are generated by BUS with subset simulation. With the posterior samples, the probability of outlier can be obtained, both block-wise and component-wise. The proposed algorithm is applied to a dataset from a clay site with some artificial outliers. The results show that the outliers can be effectively identified by the proposed outlier detection algorithm.

Acknowledgement

The authors wish to thank Prof. Dian-Qing Li and A.Prof. Zi-Jun Cao from Wuhan University for their kind helps on the 2018 TC304 Student Contest.

References

Au, S. K., and Beck, J. L. 2001. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4), 263-277.

Breunig, M. M., Kriegel, H. P., Ng, R. T., and Sander, J. 2000. LOF: identifying density-based local outliers. *ACM SIGMOD Record*, 29(2), 93-104.

Ching, J. and Phoon, K.K. 2012. Modeling parameters of structured clays as a multivariate normal distribution. *Canadian Geotechnical Journal*, 49(5), 522-545.

Rousseuw, P. J., and Leroy, A. M. 1987. *Robust regression and outlier detection*. Wiley, New York.

Straub, D., and Papaioannou, I. 2014. Bayesian updating with structural reliability methods. *Journal of Engineering Mechanics*, 141(3), 04014134.

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