

Inspection of Outliers in Multivariate Dataset by Bayesian Information Mining

J.J. Hu¹, J.Z. Hu² and T. Cheng³

¹*Department of Geotechnical Engineering, Tongji University. Email: hjj1994@tongji.edu.cn*

²*Department of Geotechnical Engineering, Tongji University. Email: tjce_hujz@foxmail.com*

³*Department of Geotechnical Engineering, Tongji University. Email: 1630299@tongji.edu.cn*

Abstract: Data from geological investigation are often sparse and subjected to many uncertainties, which brings many difficulties to geological design. Confronted with this dilemma, this TC304 student contest is about finding outliers in a limited amount of site-specific soil data. This passage adapts the Bayesian method, specifically the Gibbs Sampler (GS) method, combined with statistical hypothesis testing theory to realize the identification of outliers by comparing information between outliers and other data. Based on the algorithm and criteria, it is conventional to detect the outliers or estimate lacking data point after the outliers are removed. The application procedure and usefulness of this method will be demonstrated with the specific data contest problem.

Keywords: Bayesian analysis, multivariate soil correlation, statistical hypothesis test, data outliers, geologic site-specific investigation.

1. Introduction

Geological site-specific investigation in Geotech Engineering is necessary and valuable, because each site has its unique geological characteristics which makes geotechnical designs rely heavily on its survey. However, the key geological indicator dataset is often lacking or inaccurate due to limited conditions and human factors. To solve this problem, researches on the multivariate correlation behaviors among soil properties are greatly carried out for substitution of the missing value or validation of the soil data's intrinsic logical correctness. Ching and Phoon (2014) constructed clay transformation model based on the global database. Zhang J. and Huang H. W., et al (2012) characterized geotechnical model uncertainty by Markov Chain Monte Carlo simulation.

Under this circumstance, the TC304 student contest provided the site investigation dataset which contains several outliers and needs detecting based on some algorithms and criteria. In the following content, a data analytics method will be proposed to detecting those exception values.

2. Construction of Multi-variate Distribution

To find out the outliers, exploration on the correlation among the given data is meaningful

and worthy, which aims to construct the multivariate probability distribution for those soil properties. However, it is challenging for constructing a site-specific multivariate probability distribution. Because, if the dataset narrows down to a single site, the data points can be too sparse to construct the distribution with acceptable statistical significance. As the contest dataset provided, there are only 85(17×5) data points. It is difficult to directly establish the accurate multivariate distribution, especially some of the data are even not exactly right.

In the case where the amount and quality of data is limited, many machine learning methods such as Artificial Neural Network (ANN) and Random Forest (RF) are usually unsuitable which needs a large number of accurate data for training. Therefore, this passage adapts a Bayesian method for constructing the site-specific multivariate probability distribution that can adjust to very sparse and inaccurate site-specific data while quantifying the associated large statistical uncertainties correctly. After the distribution established, the outliers can be detected by some inspection and criteria. The investigation dataset the contest provided contains 5 column of indicators and 17 rows of values, as shown in Table 1.

Table 1. Site Investigation Dataset

LI	σ'_v (kPa)	σ'_p (kPa)	S_u^{re} (kPa)	S_u (kPa)
0.98	3.7	13.87	0.88	5.95
1.31	7.4	12.95	0.59	4.29
1.78	13.87	9.25	0.39	4.07
1.51	17.57	17.57	0.39	5
1.31	21.27	45.12	0.39	5.95
1.34	24.05	21.27	0.59	6.43
1.63	27.75	24.05	0.39	7.62
1.42	31.45	24.97	0.68	16.74
2.52	35.14	29.6	0.68	7.86
1.27	39.77	29.6	0.78	12.38
1.21	44.39	30.52	0.88	13.1
1.38	49.02	36.07	0.98	13.81
1.45	51.79	55.49	1.18	17.38
1.51	58.27	60.12	1.37	13.1
1.22	61.97	48.09	0.98	18.57
1.18	66.59	72.14	0.88	17.14
0.93	71.21	97.11	1.18	26.19

LI = liquidity index; σ'_v = vertical effective stress; σ'_p = preconsolidation stress; S_u^{re} = remolded undrained; shear strength; S_u = undrained shear strength.

The normalized data points is shown in Fig 1. It can be found that the σ'_v index has an obvious linear feature. This phenomenon is most likely because the data points are distributed with depth of the ground layer.

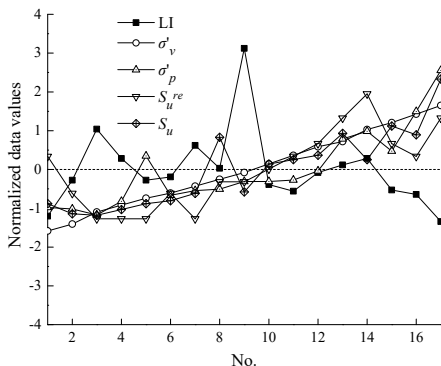


Figure 1 Normalized data points visualization

The σ'_v is vertical effective stress which is highly related to geologic conditions and the site investigation program. The other indexes may have a strong correlation with σ'_v . Let us denote that $X_1=LI$, $X_3=\ln(\sigma'_p)$, $X_4=\ln(S_u^{re})$, $X_5=\ln(S_u)$. While normal distribution is the mostly used

distribution in engineering, to simplify the problem, take this assumption that (X_1, X_3, X_4, X_5) follows multivariate normal distribution conditional on σ'_v . For σ'_v , because of its linear characteristics, it needs to be considered separately to construct its distribution. So by adopting this approach, assuming $\sigma'_{vi} = \sigma'_{v0} + \varepsilon_i$ for each σ'_v column data point, where σ'_{v0} is following discrete uniform distribution and random error term ε_i follows independent normal distribution with a zero mean and same variance. By using the maximum likelihood method, the cumulative distribution function of σ'_v , $F(\sigma'_v)$ is obtained, and then convert $F(\sigma'_v)$ to new normal variable with inverse normalized transformation. Denote the new variable as X_2 . By using Johnson distribution (Johnson, 1949; Ching and Phoon, 2014, 2018), it can be converted to normally distributed data. The calculation can be expressed as :

$$X_2 = \Phi^{-1}(F(\sigma'_v)) \quad (1)$$

After the transformation, the site-specific properties $X=(X_1, X_2, \dots, X_5)$ is then multivariate normal:

$$\begin{aligned} f(\mathbf{x} | \boldsymbol{\mu}, \mathbf{C}) &= \text{multinorm}(\mathbf{x} | \boldsymbol{\mu}, \mathbf{C}) \\ &= |\mathbf{C}|^{-\frac{1}{2}} (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \end{aligned} \quad (2)$$

where n is the dimension of the multivariate PDF ($n=5$ for this passage); $\boldsymbol{\mu}$ = the mean vector and \mathbf{C} = covariance matrix.

3. Bayesian Analysis

The Bayesian data mining approach contains two main steps. Firstly, construction of the conditional distribution $f(x_i | \mathbf{X}_0)$, where x_i is the data inspected and the \mathbf{X}_0 denotes the other data in Table 1. In the second step, simulation of the property value for each data point to update soil information. In the following section, we will explain the algorithms and analysis method for the specific contest question.

3.1 Gibbs sampler with conjugate prior model

As mentioned above, the parameters $\boldsymbol{\mu}$ and \mathbf{C} are unknown and need inferring by the existing site-specific data. Statistical theory shows that the conjugate prior distribution for $\boldsymbol{\mu}$ is multivariate normal distribution if other parameters are known, whereas the conjugate

prior distribution for \mathbf{C} is inverse-Wishart distribution. The conjugate prior distribution of $\boldsymbol{\mu}$ and \mathbf{C} should be non-informative. The prior distribution can be made non-informative by adopting large variances.

There are very few data available, it is desire to generate a large number of $(\boldsymbol{\mu}, \mathbf{C})$ samples to research the distribution nature by some statistic method. Ching and Phoon (2018) showed that it is possible to draw $(\boldsymbol{\mu}, \mathbf{C})$ samples from $f(\underline{x}|\mathbf{X}_0)$ in a statistical manner by adapting the Gibbs sampler (GS) (Geman 1984) in a conjugate prior PDF as described above. Furthermore, the contest dataset contains outliers, which makes us doubt its accuracy. Let denote the data subjected to be tested as x_i . The predicted value if x_i is left out, which is denoted as random variable, \mathbf{X}_i , can also be sampled by GS method. To sum up, the final method is to divide the parameters into a tuple, $(\boldsymbol{\mu}, \mathbf{C}, \mathbf{X}_i)$, and utilize GS method to generate the conjugate distribution. The conditional PDFs mathematical forms are:

$$\begin{aligned} \boldsymbol{\mu} &\sim f(\boldsymbol{\mu}|\mathbf{C}, \mathbf{X}_0, \mathbf{X}_i) \\ \mathbf{C} &\sim f(\mathbf{C}|\boldsymbol{\mu}, \mathbf{X}_0, \mathbf{X}_i) \\ \mathbf{X}_i &\sim f(\mathbf{X}_i|\boldsymbol{\mu}, \mathbf{C}, \mathbf{X}_0) \end{aligned} \quad (3)$$

Statistical theory shows that the first two posterior PDFs are still same with prior distribution forms: $f(\boldsymbol{\mu}|\mathbf{C}, \mathbf{X}_0, \mathbf{X}_i)$ is still multivariate normal as stated above, and $f(\mathbf{C}|\boldsymbol{\mu}, \mathbf{X}_0, \mathbf{X}_i)$ is still inverse-Wishart. Moreover, $f(\mathbf{X}_i|\boldsymbol{\mu}, \mathbf{C}, \mathbf{X}_0)$ is also multivariate normal due to the assumed multivariate normality. Thus, with the aforementioned analysis, the GS algorithm can be executed expediently, because all PDFs in Eq. 2 are able to be sampled easily. The variables in tuple $(\boldsymbol{\mu}, \mathbf{C}, \mathbf{X}_i)$ are connected and affected by each other. For GS realization, the sampler begins with an initial random value $(\boldsymbol{\mu}_0, \mathbf{C}_0, \mathbf{X}_{i0})$, and then continuously perform drawing samples $(\boldsymbol{\mu}_n, \mathbf{C}_n, \mathbf{X}_{in})$ (n , execution times=1,2,..., N) based on the latest generated parameter values with the conditional PDFs in Eq. 3. After constant sampling, or called the burn-in period, the tuple $(\boldsymbol{\mu}, \mathbf{C}, \mathbf{X}_i)$ distribution is constructed, which is convincing that the distribution reflects the site soil information and quantifies the specific uncertainty, for whose samples are drawn from the statistics of the existing data points.

For specific contest question, each data value needs examining and sampling with the above method as the outliers are unknown. By

observing the LI column in Table 1, the value in the 9th row (2.52) is too large relative to other column numbers which is initially subjectively judged to be an outlier. Take this assumption for illustration, the test value x_i is 2.52. With GS algorithm, it is convenient to establish the $(\boldsymbol{\mu}, \mathbf{C}, \mathbf{X}_i)$ distribution. Fig 2 shows the $\text{LI}|\mathbf{X}_0$ sample trace which reveals the samples tend to converge at the end.

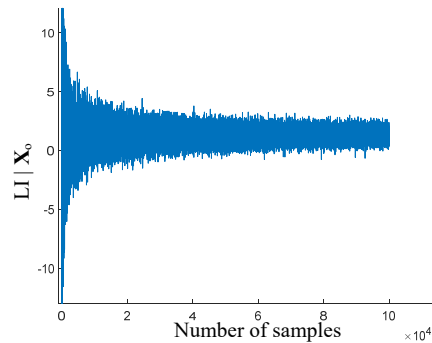


Figure 2 Trace of $\text{LI}|\mathbf{X}_0$ Samples

3.2 Data test for outliers

After constructing of $f(\mathbf{X}_i|\mathbf{X}_0)$, simulation of the \mathbf{X}_i is the next job to update soil information and verify whether the original value or variable follows the posterior distribution at a given acceptable significance. Here a significance level $\alpha = 0.05$ is adopted. The Student's t-test is applied for inspection in this paper, which is widely used for univariate distribution test in statistics

Let continue with the previous example: The \mathbf{X}_i could simulate from the samples we have drawn, because of the final convergence the Fig.3.1 shows. In our updating process, the last quarter of samples are taken to establish the prior PDFs with Eq.4. Fig. 3 shows the histogram of the \mathbf{X}_i when the original value is 2.52. Just what the figure shows, in the case when the significance level is 0.05, the original value falls outside the confidence interval. Therefore, A preliminary judgment can be made that the original value 2.52 may be an outlier.

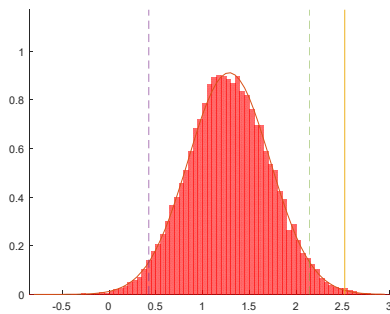


Figure 3 Histogram of samples ($x_i = 2.52$)
 (the solid line on the right reflects the original value 2.52 location, the two dashed lines indicate the confidence interval)

According to the method described in the previous example, all the data points can be tested respectively. Fig. 4 reveals all of the data points examination result. From the chart, each histogram position stands for the location of data in Table 1 where the red histograms indicate the original value is most likely an outlier. The upper and lower boundaries of 95% confident interval are plotted with dashed line and the solid line stands for the origin value in Table 1.

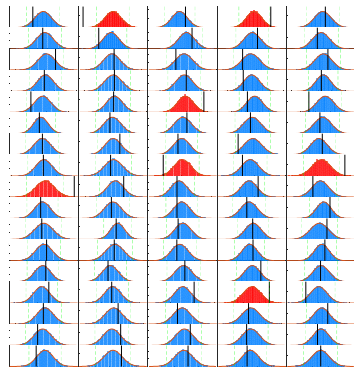


Figure 4 Test result for each datum

Through the above method and criteria, 7 outliers could be initially detected. In order to express fairly and simply, the outliers are recorded as a tuple (origin value, row, column, soil index), for example, the outlier (2.52, 9, 1, LI) is value 2.52 at 9th row and 1st column in Table 1, which is a LI soil index. Then, the outliers are:
 (2.52, 9, 1, LI), (3.7, 1, 2, σ'_v), (45.12, 5, 3, σ'_p),
 (24.97, 8, 3, σ'_p), (0.88, 1, 4, S_u^{re}),
 (1.37, 14, 4, S_u^{re}), (16.74, 8, 5, S_u).

4. Further inspection with several suspected outliers

From above analysis, the outliers can be identified preliminarily. However, it cannot be ignored that the outliers we have detected are inspected based on the distribution sampling from the other data, X_0 , including other outliers which are inaccurate. Some data points could be misjudged as outliers in some extreme conditions. Fig. 5 shows that the data point 1.37 is determined as outlier (1.37, 14, 4, S_u^{re}), while it is very close to the acceptable interval, where the solid line and the right dash line are most coincidence.

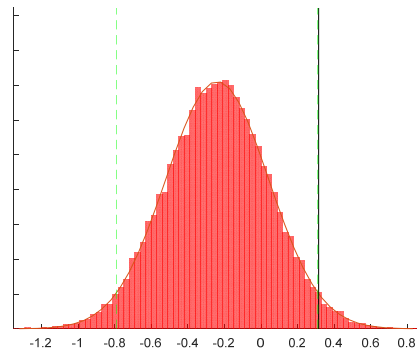


Figure 5 Histogram of samples ($x_i = 1.37$)

For this reason, it is necessary to exclude these outliers from the dataset separately and inspect them with the methods described in Chapter 3. The inspection before is univariate distribution test for each testing data point X_i , where the Student's t-distribution can complete this task.

While combining the 7 suspected outliers altogether and inspect them with Student's t test, there are some "outliers" accepting the null hypothesis and falling within the acceptable intervals, which means they are "non-outliers" under this circumstance. Fig. 6 shows this phenomenon.

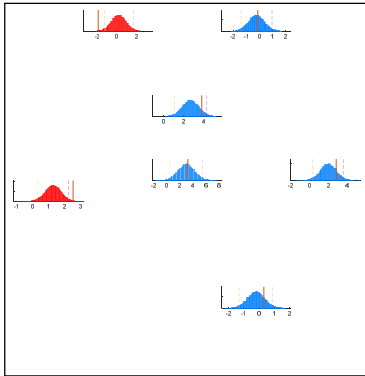


Figure 6 Test result for 7 suspected outliers

So, how to determine the outliers and choose the combination? Our criterion is to find as many outliers as possible, and at the same time, there is no “non-outliers” in such combination that reject the null hypothesis. This is to minimize the probability of Type II error as small as possible. Through constant experimentation and enumeration, it meets the requirements when these outliers (2.52, 9, 1, LI), (3.7, 1, 2, σ'_v), (45.12, 5, 3, σ'_p) are combined. Fig. 7 shows this situation.

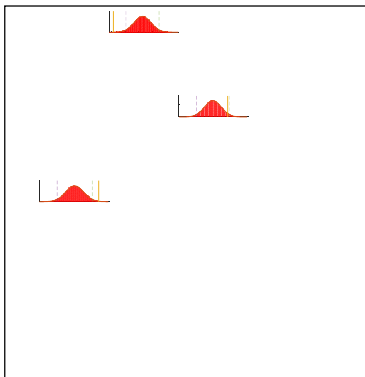


Figure 7 Test result for 3 suspected outliers

Finally, we determine these 3 data points as outliers based on the methods and criteria we put forward. While combining these outliers together, the inspection is a multivariate test problem. Hotelling's T-squared test is a generalization of Student's t-statistic that is used in multivariate hypothesis testing which can solve this problem. The Hotelling's T-squared test can once inspect whether the whole vector \underline{X}_i is from the same multivariate normal

distribution or not. The test P value of Hotelling's T-squared test for these is 6.25×10^{-5} , which is rejected under the given significance level.

The 3 final suspected outliers are also filled with grey color in Table 1. As expected, the LI in the 9th row and σ'_p in the 5th row is larger than usual and σ'_v in the 1st row is smaller. There is a trade-off subjected to the correlation between soil properties. For example, there may be a trade off between (2.52, 9, 1, LI) and (16.74, 8, 5, S_u) for higher liquidity index usually leads to low strength. They are both suspected outlier at the beginning. Higher liquidity index generally leads to lower strength. If (2.52, 9, 1, LI) is accepted, the rejection of (16.74, 8, 5, S_u) can be more significant and vice versa. The result comes from the largest significance or smallest p value, which may lead to type II error for the accepted one.

It should be noted that this method just lead to largest difference between suspected outliers and other data based on statistical inference. The final decision also calls for engineering experience due to some unclear random factors, e.g. determination of significance level.

5. Conclusion and Discussion

This passage adapts the Bayesian method, combined with statistical hypothesis testing theory, to find outliers in site investigation when the site-specific data are few and inaccurate. Through the method and criterion we propose, we establish the multivariate correlation among soil properties and detect the outliers to a certain extent. While this discrimination is largely based on the selection of significance level and the criteria of accepting or rejecting, it is still a valuable method to learn from for geological survey and design when choosing the appropriate standard.

Acknowledgement

The authors would like to thank Prof. Huang and Prof. Zhang from Tongji University for their kind support and patient instruction.

References

- Ching, J. and Phoon, K.K. 2014 Transformations and Correlations among Some Parameters of Clays—the Global Database, *Canadian Geotechnical Journal*, 51(6), 663-685,
- Ching, J. and Phoon, K.K. 2018. Bayesian data mining for a generic geotechnical database. In *Proceedings of the 6th Intl. Symposium on Reliability Engineering and Risk Management*. Singapore ,31 May - 1 June 2018, pp. 17-24
- Ching, J. and Phoon, K.K. 2018. Constructing multivariate probability distribution for soil properties based on site-specific data. In *Proceedings of the 6th Intl. Symposium on Reliability Engineering and Risk Management*. Singapore ,31 May – 1 June 2018, pp. 255-260
- Geman, S. and Geman, D., 1984. Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721-741.
- Johnson, N. L., 1949, Systems of Frequency Curves Generated by Methods of Translation, *Biometrika*, 36, 149-176.
- Zhang J, Tang W H, Zhang L M, et al. 2012, Characterising geotechnical model uncertainty by hybrid Markov Chain Monte Carlo simulation. *Computers and Geotechnics*, 43: 26-36.