Bandwidth Selection of Normal Information Spread Estimation Method for Geotechnical Parameter Probability Distribution

KONG Lingqi

1School of Materials and Environmental Engineering, Chengdu Technological University, Chengdu 610031, China
*Kong lingqi’s e-mail: k990418@163.com

Abstract. Normal information spread estimation method (NISEM) is an effective method to infer the geotechnical parameter probability density function. However, it is found that the optimal bandwidth calculated by the existing method is not consistent with the actual situation in engineering application. In order to determine the matching bandwidth of NISEM, the histogram grouping number and group distance according to the sample quantity and range are first determined in this paper. Then the bandwidth h of NISEM is calculated, taking h as the benchmark, 0.1h as the step to change bandwidth. And combining with judgment coefficient, the optimal matching bandwidth of NISEM is finally obtained. The validity and rationality of the proposed method are verified by geotechnical test data of two groups of large samples and four groups of small samples in practical engineering. The results show that: under the condition of limited geotechnical samples, the best matching bandwidth is related to the number of histogram groups, which is relative. And the bandwidth of each group of test samples is different. The research work in this paper perfects the application of NISEM in inferring geotechnical parameter probability density function.

1. Introduction

The probability distribution model of geotechnical parameters directly affects geotechnical engineering reliability index calculation results in geotechnical engineering risk assessment and reliability analysis. Therefore, the research on geotechnical parameter probability distribution has always been a basic research work, which has extremely important value and significance [1-3]. A lot of work about geotechnical parameter probability distribution has been done both in China and overseas [4-7]. Normal, lognormal and beta distribution model-based on parameter estimation are adopted to fit probability distribution model of geotechnical parameters [8-9] in early researches, and it is found that the method-based on parameter estimation results in relatively large fitting error with the deepening of research. In particular, the mismatch problem exists between definition interval and actual geotechnical parameter distribution interval in normal and lognormal distribution [10]. Then the maximum entropy method, orthogonal polynomial method and other methods-based on non-parametric estimation are proposed to fit geotechnical parameter probability distribution [1-3, 10-12]. The NISEM is finally determined as the relatively optimal inference method to infer geotechnical parameter probability distribution in [13], after advantages and disadvantages are systematically investigated through studying that adapt classical distribution, maximum entropy method, general polynomial method, orthogonal polynomial method and NISEM to infer geotechnical parameter probability distribution. The effectiveness of NISEM is further...
proved in the probability distribution inference of large sample geotechnical parameters in [1]. Although the optimal bandwidth is given in [1], but the matching relationship between bandwidth selection and frequency histogram is neglected. Moreover, as for geotechnical parameter test samples, the test data of geotechnical parameters are limited due to the limitation of economic and technical conditions [14], it is difficult to satisfy the condition of obtaining the optimal bandwidth value. Besides, it is difficult to determine the bandwidth of NISEM, because bandwidth value will affect fitting curve shape, then affect fitting accuracy [15].

Summarizing existing researches, it can be found that the fitting methods of geotechnical parameter probability density function are in essence probability distribution fitting based on histogram. Whether the fitting methods-based on parameter estimation or methods-based on nonparametric estimation, they are all firstly carried out based on histograms which come from test samples. Histogram drawing depends on interval division, and different interval divisions directly affect the fitting curve shape and fitting error. At present, it is difficult to ensure the interval division rationality for the number of histograms and interval division are mainly determined by researchers subjectively, in the inferring process of geotechnical parameter probability distribution model. Although the optimal bandwidth calculation formula is given in [1], it is not optimal in fact because it neglects the matching relationship between bandwidth selection and histogram, so it is relative optimal bandwidth.

The contributions of this paper can be outlined as follows: Sample grouping number and group distance are obtained according to the number of samples and the range, to make the histogram closer to actual probability distribution of samples; The “method for determining optimal matching bandwidth” of NISEM is proposed on the basis of appropriate histogram grouping; At last, validity and rationality of proposed determining method are verified by engineering test samples, and the application of NISEM in geotechnical parameter probability density function fitting process is improved.

2. Existing bandwidth determination method

2.1 Probability density function of NISEM

In order to avoid adverse influence of interval partition on probability density function curve shape, the NISEM considers certain neighborhood centered on anyone point $x$, then counts the number of sample points in this neighborhood. The probability density function value at point $x$ is determined by the number of sample points in the region.

Assuming that $X_1, X_2, \ldots, X_n$ are samples from the population, then the probability density function of samples is shown in (1) [1, 16-18]:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{\sqrt{2\pi nh}} \sum_{i=1}^{n} e^{-\frac{(x - X_i)^2}{2\sigma^2}}$$

In (1), $K(\cdot)$ is kernel function, $h$ is bandwidth, and $n$ is the number of samples.

2.2 Existing bandwidth determination method

Bandwidth is an important index when adopts NISEM to infer geotechnical parameter probability density function. The optimal bandwidth is obtained by solving the minimum value of mean integrated squared error (MISE) function in [1], corresponding calculation formula is shown in (2):

$$\text{MISE} \left( \hat{f}_n \right) = E \left[ \int \left( \hat{f}_n(x) - f(x) \right)^2 dx \right]$$

In (2), $f(x)$ denotes actual probability density of sample population. $K(x)$ satisfies following four conditions:

- $K(x)$ gets a value in the interval [-1,1], and it is symmetry; $\int K(x) dx = 1$; $\int xK(x) dx = 0$;
- $\int x^2 K(x) dx = \sigma_x^2 > 0$, $\sigma_x$ is standard deviation.

Then
\[
MISE(\hat{f}_h) \approx \frac{1}{4} \sigma_i^4 h^4 \int \left[ f''(x) \right]^2 dx + \frac{1}{nh} \int \left[ K(x) \right]^3 dx \tag{3}
\]

Solve the formula \( \min_h MISE(\hat{f}_h) \), then optimal bandwidth is obtained as shown in (4).

\[
\hat{h} = \left\{ \left[ \frac{\int \left[ K(x) \right]^3 dx}{\sigma_i^4 \int \left[ f''(x) \right]^2 dx} \right]^{\frac{1}{3}} \right\}^{\frac{1}{3}} \tag{4}
\]

It is worth noting that one most important precondition is essential for obtaining (4), that is, only if \( h \to 0, \ nh \to +\infty \) can make left and right sides of equal sign in (4) is approximately equal \[16\]. But the precondition are difficult to be satisfied for geotechnical parameter test quantity is not large enough, due to the limitation of economic and technical conditions \[14\]. Therefore, for geotechnical parameter test samples, the calculated bandwidth from (4) is not optimal when utilizes NISEM to fit the geotechnical parameter probability density function.

2.2.1 The influence of histogram grouping on fitting curve

Test quantity of Kolmogorov-Smirnov test is the maximum deviation value in the whole value range, the validity of a certain distribution fitting can be evaluated by Kolmogorov-Smirnov test under certain significance level. But Kolmogorov-Smirnov test cannot provide absolute information whether a certain distribution fitting is good or not \[19, 20\]. In order to evaluate the advantages and disadvantages of fitting method, a dimensionless judgment coefficient is introduced to compare and analyze the fitting estimated value with the actual value, and the coefficient is an index for visually evaluating the fitting effect \[21-23\]. The judgment coefficient is defined as follows:

\[
R^2 = \sum_{i=1}^{n} \left( Y_i - \overline{y} \right)^2 / \left( \sum_{i=1}^{n} (Y_i - \overline{y})^2 + \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2 \right) \tag{5}
\]

In (5), \( y_i \) is measured value, \( \overline{y} \) is the average of measured values, \( \overline{Y}_i \) is fitting estimated value. According to (5), the smaller the error between \( y_i \) and \( \overline{y} \) is, the closer the judgment coefficient is to 1, the better the fitting effect will be.

Select 81 groups of measured data of geotechnical parameter test samples from \[1\], the measured cohesion value \( c \) is \[1\]: 30, 27, 19, 11, 13, 33, 26, 12, 32, 15, 21, 7, 28, 10, 9, 29, 22, 14, 32, 8, 34, 13, 21, 10, 15, 25, 28, 15, 38, 26, 13, 15, 27, 25, 16, 17, 20, 20, 30, 18, 15, 20, 7, 17, 20, 19, 24, 16, 35, 12, 30, 19, 13, 16, 14, 15, 43, 35, 24, 17, 25, 13, 30, 17, 25, 19, 13, 35, 13, 18, 30, 20, 30, 12, 29, 15, 30, 11, 34, 25, 33, their unit is kPa.

For cohesion samples \( c \), histogram grouping numbers are set as 5, 10 and 20 respectively. The bandwidth \( h \) is calculated according to (4) and it is 0.1024 after calculation. The obtained fitting curves via NISEM are shown in Fig.1-Fig.3. In order to eliminate the limitation from dimension and magnitude order of each sample, the range normalization transformation is carried out \[16\] on geotechnical parameter test samples in this paper, namely

\[
x_{i}^\prime = (x_i - \min_{1 \leq i \leq 10} x_i) / (\max_{1 \leq i \leq 10} x_i - \min_{1 \leq i \leq 10} x_i)
\]

Where \( \max_{1 \leq i \leq 10} x_i \) and \( \min_{1 \leq i \leq 10} x_i \) denote maximum and minimum of each group test sample.

![Fig.1 The number of group is 5](image1)

![Fig.2 The number of group is 10](image2)
Fig. 3 The number of group is 20

For cohesion test samples $c$, the bandwidth is selected as $h$. The judgment coefficient is 0.9692 when the number of histogram group is 5. The judgment coefficient is 0.9502 when the number of histogram group is 10. The judgment coefficient is 0.7410 when the number of histogram group is 20.

According to above analysis, it can be seen that: different grouping numbers of histograms reflect the volatility and discreteness in different degrees of samples, and corresponding bandwidth values should also be different. When the fitting bandwidth $h$ of NISEM is determined according to (4), the fitting accuracy is different with different histogram grouping numbers. So matching relationship exists between NISEM fitting bandwidth and histogram grouping number, and the bandwidth width is relative.

It can be clearly seen from Fig. 1 to Fig. 3 that: the NISEM matching bandwidth should be different when the number of histogram group is different. The larger the histogram grouping number is, the probability distribution shows multiple peak points, but these local characteristics cannot be reflected by using a larger bandwidth. Therefore, the smaller the histogram grouping number is, the larger the matching bandwidth should be. The larger the histogram grouping number is, the smaller the matching bandwidth should be.

2.2.2 Influence of bandwidth on geotechnical parameter probability distribution curve

In this part, analyze the influence of different bandwidth on fitting curve shape by NISEM, through determining the histogram grouping number and changing the bandwidth.

For the cohesion test samples $c$ [1], when histogram grouping number is 10, the comparison of NISEM fitting curves under different bandwidth conditions are shown in Fig. 4. The fitting effect judgment coefficient when bandwidth changes is shown in Table 1.

![Fig. 4 Fitting curves of cohesion samples under different bandwidth conditions (1)](image)

**Table 1 Judgment coefficient of cohesion sample**

| Bandwidth | $h$ | $0.3h$ | $0.4h$ | $0.5h$ |
|-----------|-----|--------|--------|--------|
| Judgment coefficient | 0.7134 | 0.9710 | **0.9832** | 0.9830 |

According to Fig. 4 and Table 1, the histogram grouping number is taken as 10, the judgment coefficient is 0.7134 when adopts bandwidth calculated by (4) to fit the probability density function of cohesion test samples $c$. The judgment coefficient is 0.9830 when bandwidth is reduced to $0.4h$, the fitting error decreases, fitting effect is better than the fitting curve with bandwidth $h$.

For cohesion test samples $c$, histogram grouping number is taken as 20, and the fitting curves of NISEM under different bandwidth conditions are shown in Fig. 5. The fitting effect judgment coefficients under different bandwidth conditions are shown in Table 2.
According to Fig.5 and Table 2, histogram grouping number is 20, the judgment coefficient is 0.7410 when adopts bandwidth calculated by (4) to fit the probability density function of cohesion test samples. The judgment coefficient is 0.9388 when bandwidth is reduced to 0.2 \( h \), the fitting error decreases, fitting effect is better than the fitting curve with bandwidth \( h \).

To sum up, both the direct comparison of fitting curves and the quantitative test of judgment coefficient show that \( h \) is not the optimal bandwidth.

3. Matching bandwidth determination method

3.1 Appropriate histogram grouping

According to above analysis, the histogram grouping number should be determined first when adopts NISEM to infer geotechnical parameter probability distribution. The smaller the number of histogram group is, the smoother the probability density function curve is, but fluctuation characteristics of geotechnical parameter test samples cannot be reflected. The larger the histogram grouping number is, the more tortuous the probability density function curve is, and it can show fluctuation and local characteristics of geotechnical parameter test samples, but the more complex the function form of probability distribution fitting curve will be. So appropriate histogram grouping number is the premise of adopting NISEM to fit geotechnical parameter probability distribution.

According to the theory about probability and mathematical statistics, histogram approaches the probability density function of samples with high probability, which is a method to minimize fitting deviation. However, the size and number of histogram subintervals directly affect the shape and fitting accuracy of probability density function curve [19]. In order to make the histogram approach the probability density function with high probability, the number of subintervals should be selected reasonably and it is usually obtained by attempt [19].

In probability and mathematical statistics theory, for random variables which belongs to normal distribution population, the best matching relationship exists between histogram subinterval division and the number of samples. The number of groups is taken as \( m = 1.87(n-1)^{2/5} \) [19]. However, a great deal of research work has shown that the geotechnical parameter test samples show great discreteness, and they belong to skew distribution. Therefore, the histogram grouping number is set by (6) for random variable population which belongs to non-normal distribution in engineering when the number of samples is greater than 50.

\[
m = 1.52(n-1)^{2/5}
\]  

In (6), \( n \) is the number of samples.

Grouping interval is calculated by (7):

\[
Ax = R / m
\]

In (7), \( R \) is range.

In order to avoid the cross-group problem when grouping the measured values located at the boundary point, the effective number of group boundary value should be one digit more than the original.
measured value, and the effective number value of this one digit should be taken as 0.5 [24]. The average of upper and lower group boundary value is called group mid-value, which can be viewed as the representative of measured values of this group. In order to make sure that the group contains the minimum value $z_{\text{min}}$ and the maximum value $z_{\text{max}}$, the lower limit of histogram abscissa should be smaller than $z_{\text{min}}$ while the upper limit should be larger than $z_{\text{max}}$.

3.2 Matching bandwidth determination method
According to the simulation analysis on geotechnical parameter test samples, the bandwidth $h$ calculated by (4) is suitable for geotechnical parameter test samples with less volatility, but a smaller bandwidth is more suitable for test samples with greater volatility. Meanwhile, the physical and mechanical parameters of geotechnical are characterized by great fluctuation and strong discreteness. So fixed bandwidth calculation formula is difficult to meet the requirements of various geotechnical parameter test samples.

Therefore, aim at inferring the probability distribution of geotechnical parameter test samples by NISEM, this paper puts forward a bandwidth selection method. The steps that utilizing NISEM to infer geotechnical parameter probability density function are summarized as follows:

1. Input sample data, calculate the grouping number and group distance of sample histogram according to (6) and (7), then draw sample histogram. Histogram grouping should be selected tentatively when the number of samples is small.

2. Calculate the bandwidth $h$ of NISEM according to (4), on this basis of $h$, take 0.1$h$ as step size to change the bandwidth. View the maximum judgment coefficient as selection rule, and then determine the "optimal matching bandwidth" of NISEM for geotechnical parameter test.

3. Draw the probability density function curve of test samples fitted by NISEM, then output the result.

3.3 Case verification
The effectiveness of proposed method is verified through two large samples of cohesion $c$ and internal friction angle in section 1.2.2 and it’s not repeated here.

In order to verify the universality of this method, four groups of small sample geotechnical parameters are selected to illustrate it. Liquid limit sample, the capacity is 25: 11.86, 14.04, 13.5, 10.2, 11.31, 12.95, 12.41, 13.5, 12.95, 15.64, 16.17, 17.22, 20.81, 20.81, 25.64, 22.29, 24.70, 26.1, 22.78. Liquidity index sample [25], the capacity is 23: 0.59, 0.62, 0.65, 0.74, 0.77, 0.78, 0.85, 0.87, 0.91, 0.92, 0.92, 0.93, 0.95, 0.96, 0.98, 0.99, 1.03, 1.09, 1.1, 1.11, 1.15.

Due to the limitation of paper page quantity, only the probability density function curve comparison diagram via NISEM based on two kinds of bandwidth determination method is given from Fig.6 to Fig.7. Specifically, they are the optimal matching bandwidth determined by the proposed method in this paper and the bandwidth $h$. Statistical results of judgment coefficient are shown in Table 3.

| Sample          | Internal friction angle | Liquidity index |
|-----------------|------------------------|-----------------|
| $h$             | 0.5336                 | 0.7069          |
| Matching bandwidth | 0.9637              | 0.9566          |

Fig.6 Fitting curve of internal friction angle sample  
Fig.7 Fitting curve of liquid limit index sample
According to Fig.6-Fig.7 and Table3 in NISEM application process, the probability density function obtained with the optimal matching bandwidth method proposed in this paper is better than the probability density function obtained under the bandwidth $h$ condition. This can be seen from both comparison of fitting curves and quantitative test of judgment coefficient. The correctness and applicability of the proposed bandwidth selection method in NISEM are illustrated.

4. Conclusion
In this paper, the application of NISEM in fitting geotechnical parameter probability density function is improved on the basis of existing research work. Several main conclusions are obtained as follows:

(1) Appropriate sample histogram grouping is the premise of inferring geotechnical parameter probability density function by NISEM. The larger the histogram grouping number of test samples is, the smaller the matching bandwidth should be adopted. The smaller the histogram grouping number of test samples is, the larger the matching bandwidth should be adopted.

(2) After the histogram groupings number is determined, the "optimal matching bandwidth" method proposed in this paper is simple and effective. Which is based on the bandwidth $h$ and takes $0.1h$ as the step size to change bandwidth, then determine "optimal matching bandwidth" via judging the maximum judgment coefficient. This bandwidth determination method is applicable to both large and small samples of geotechnical parameter test in engineering.

References
[1] Zhu Yuanzhen, Li Xibing, Gong Fengqiang. Inference of probability distribution of rock and soil parameters by normal information diffusion [J]. Rock and Soil Mechanics, 2015, 36(11):3275-3282.
[2] GONG Feng-qiang, LI Xi-bing. Structural Reliability Analysis Based on Legendre Orthogonal Polynomial Approximation [J]. Engineering Mechanics, 2008(06):225-229.
[3] CHEN Wang-wang, LI Dian-qing, TANG Xiao-song, CAO Zi-jun. Probability distribution of shear strength parameters using maximum[J]. Rock and Soil Mechanics, 2018,39(04):1469-1478.
[4] HUANG Tian-lang, GONG Feng-qiang, WU Ting-yu. Orthogonal polynomial inference method of the probability distribution for small samples of geotechnical parameters[J]. ENGINEERING MECHANICS,2017,34(08):85-95.
[5] Li Dianqing, Tang Xiaosong, Phoon Kok-kwang, et al. Bivariate simulation using copula and its application to probabilistic pile settlement analysis [J]. International Journal for Numerical and Analytical Methods in Geomechanics, 2013, 37(6): 597—617.
[6] Phoon Kok-kwang, Ching Jianye. Risk and reliability in geotechnical Engineering [M]. London and New York: Taylor & Francis, 2015: 339—433.
[7] Gong Fengqiang, Huang Tianlang. Sample size effect on the probability distribution fitting accuracy of random variable by using normal diffusion estimation method-compared with normal distribution [C] Proceedings of the 6th Asian-Pacific Symposium on Structural Reliability and its Applications. Shanghai: Tongji University Press, 2016: 199—204.
[8] Meng Qingshan, Lei Xuewen. Statistical methods and application to engineering of geotechnical parameters [J].Journal of Wuhan Yejin University of Science and Technology, 1999, 22(4): 414—417.
[9] LU Yan-er,LIU Yong. Beta Probability Distribution of Shear Stress Parameters of Weakness Interlayers in Rockmass[J]. Chinese Journal of Underground Space and Engineering,2014,10(06):1250-1256.
[10] GONG Feng-qiang, HUANG Tian-lang, LI Xi-bing. Inference method for optimal probability distribution function of shear strength parameters in geotechnical engineering [J]. Chinese Journal of Geotechnical Engineering, 2016,38(S2):204-209.
[11] GONG Feng-qiang, HOU Shang-qian, YAN Xiao-ming. Probability model deduction method of Mohr-Coulomb criteria parameters based on normal information diffusion principle[J].
Chinese Journal of Rock Mechanics and Engineering, 2013, 32(11): 2225—2234.

[12] HUANG Tian-lang, GONG Feng-qiang, WU Ting-yu. Orthogonal Polynomial Inference Method for Probability Distribution of Small Sample Geotechnical Parameters [J]. Engineering Mechanics, 2017,34(08):85-95.

[13] Gong fengqiang, Huang tianlang, Li xibing. study on inference method and criterion of optimal probability distribution of geotechnical parameters [J]. Chinese journal of rock mechanics and engineering, 2016,35(12):2452-2460.

[14] Lei Zhang. Study on two-dimensional distribution model of rock and soil parameters based on Copula function [D]. Wuhan University, 2017.

[15] Jiang Shuihua, Li Dianqing. Bayesian method for updating slope reliability. [M]. Beijing: Science Press, 2019.

[16] Xie Zhonghua. Statistical analysis and application of MATLAB [M]. Beijing: Beihang University Press, 2010.

[17] Zhang Dongdong. Research on finite element model validation method based on Gaussian process response surface and kernel density estimation [D]. Nanjing University of Aeronautics and Astronautics, 2012.

[18] Wen Shijun. Research on Kernel Non-method and Its Application in Probability Density Estimation [J]. Shandong Industrial Technology, 2013(10):163-164+197.

[19] Chen lizhou, He Xiaofeng, Weng haishan. design of engineering variable optimization method-principle and application [M]. Beijing: science press, 1997.

[20] Deng Bo. Statistical processing method for analyzing test data [M]. Beijing: Tsinghua University Publishing House, 1995.

[21] Yan wang, Sui Silian, Wang Aiqing. Mathematical Statistics and MATLAB Engineering Data Analysis [M]. Beijing: Tsinghua University Publishing House, 2006.

[22] Translated by Sun Fang. the concept of probability in engineering planning and design [M]. Beijing: smelting industry press, 1991.

[23] XUChao, YANG Lin-de. Test of goodness of fit of random variables and Bayesian estimation of distribution parameters[J]. Journal of Tongji University, 1998, (3): 340－344.

[24] Lynn Yang. Analysis of generator overvoltage and its influence on insulation aging [D]. Southwest Jiaotong University, 2008.

[25] DENG Jian, LI Xi-bing, GU De-sheng. probability distribution of rock mechanics parameters by using maximum entropy method [J]. Chines Journal of Rock Mechanics and Engineering, 2004,23(13):2177-2181.