Bayesian Learning of Gaussian Mixture Model of Geotechnical Data

Qin-Xuan Deng[1], Jian He[2], Zi-Jun Cao[1], and Dian-Qing Li[1]

[1] State Key Laboratory of Water Resources and Hydropower Engineering Science, Institute of Engineering Risk and Disaster Prevention, Wuhan University, 8 Donghu South Road, Wuhan 430072, P. R. China.
[2] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.
E-mail: january@whu.edu.cn; jhebl@connect.ust.hk; zijuncao@whu.edu.cn (corresponding author); dianqing@whu.edu.cn

Abstract: Uncertainties in geotechnical parameters are unavoidable and can be quantified through probabilistic models. However, determining probabilistic model is a fundamentally ill-posed problem in the sense that there are infinitely many probability models that can generate the same set of geotechnical test data. Convention in geotechnical reliability and risk often assumes simple probability distributions, such as Gaussian and lognormal distributions, for mathematical conveniences, neither of which can be universally applicable in terms of goodness-of-fit with geotechnical data acquired at different sites. In contrast, Gaussian mixture model (GMM) provides great flexibility in fitting geotechnical test data that might vary in a wide range while inheriting many mathematical conveniences of the Gaussian distribution. However, GMM can easily overfit the test data. This paper develops a Bayesian learning method that combines Bayesian model class selection with GMM to identify the probabilistic model of geotechnical data. As the number of component Gaussian PDF increases, the number of GMM parameters increases, leading to a profound computational difficulty in Bayesian updating. Such a computational difficulty is tackled using Bayesian updating with structural reliability methods (BUS) in this study. The proposed approach is applied to analyzing geotechnical data in 304dB, geotechnical databases compiled by International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE) TC304.

Keywords: Gaussian mixture model; Bayesian learning; model selection; BUS; geotechnical data.

1 Introduction

Uncertainties are unavoidable in geotechnical engineering, and they can arise in loads, geological site interpretation, geotechnical parameters, calculation models, etc. Among these geotechnical-related uncertainties, the uncertainty in geotechnical parameters is one of the most important uncertainty sources affecting decision making in geotechnical designs and analyses. How to properly model and deal rationally with the uncertainty in geotechnical parameters is an intriguing issue in geotechnical engineering.

Probability and statistics have been used in previous studies to model the uncertainty in geotechnical parameters. By this means, soil parameters were often modelled using simple probability distributions, such as Gaussian and lognormal distributions (e.g., Lumb 1966; Lacasse and Nadim 1996). These simple probabilistic models of soil parameters are attractive due to the mathematical conveniences provided by them for subsequent reliability analysis and risk assessment. However, it shall be noted that determining probabilistic models of soil parameters from a limited number of test data is a fundamentally ill-posed problem in the sense that there are infinitely many probability models that can generate the same set of geotechnical test data. Performance of the simple probabilistic models (e.g., Gaussian and lognormal random variables) can vary significantly from one site to another in terms of goodness-of-fit with geotechnical data acquired at different sites.

In comparison with simple probabilistic models, Gaussian mixture model (GMM) (McLachlan and Peel 2000; Bishop 2006; Wang et al., 2015) provides great flexibility in fitting geotechnical data varying in a wide range while inheriting many mathematical conveniences of the Gaussian distribution. However, GMM can easily overfit the test data, particularly when the number of data is limited. This necessitates a proper choice of the number of Gaussian components of GMM for trade-off between the goodness-of-fit and predictability of GMM.

This paper develops a Bayesian learning method that combines Bayesian model class selection with GMM to identify a proper probabilistic model of geotechnical data. It starts with a brief introduction of Gaussian mixture model, followed by development of the proposed method. Then, Bayesian updating with structural reliability methods (BUS) using Subset Simulation (Straub and Papaioannou 2015) is used to evaluate the model evidence for model class selection and to generate posterior samples of model parameters for uncertainty quantification. The proposed method is illustrated using real data, specifically F-CLAY/7/216, in 304dB-

Proceedings of the 7th International Symposium on Geotechnical Safety and Risk (ISGSR)
Editors: Jianye Ching, Dian-Qing Li and Jie Zhang

Published by Research Publishing, Singapore.
ISBN: 978-981-11-2725-0; doi:10.3850/978-981-11-2725-0_MS2-1-cd

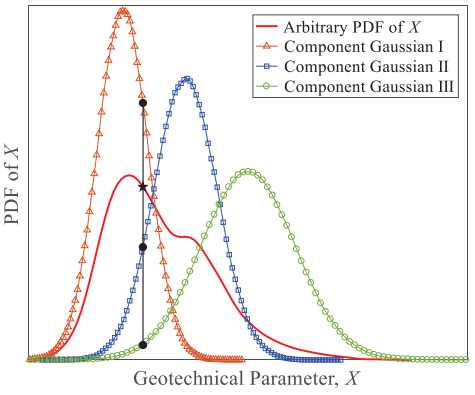


Figure 1. Illustration of Gaussian mixture model .

geotechnical databases compiled by International Society for Soil Mechanics and Geotechnical Engineering (ISSMGE) TC304 (http://140.112.12.21/issmge/tc304.htm).

2 Gaussian Mixture Model

Mixture model provides a convenient semi-parametric framework for representing unknown distribution (McLachlan and Peel 2000). In the context of mixture model, the probability density function (PDF) of a random variable X is expressed as a weighted linear combination of a number, K, of component PDFs. Consider, for example, taking the Gaussian PDF as the component PDF of the mixture model. This results in a Gaussian mixture model (GMM) of X, and its PDF is written as (e.g., McLachlan and Peel, 2000; Bishop, 2006; Wang et al., 2015)

$$p(X \mid \boldsymbol{\omega}_K) = \sum_{k=1}^{K} \pi_k N(X \mid \mu_k, \sigma_k) \tag{1}$$

where $N(X \mid \mu_k, \sigma_k)$ = the k-th component Gaussian PDF with a mean value μ_k and a standard deviation σ_k; π_k = component weighting coefficient of the k-th component, which satisfies

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{for } k = 1, 2, ..., K$$
$$0 < \pi_k < 1 \tag{2}$$

Equation (1) can be used to represent an arbitrary PDF of X by adjusting the number of component PDFs and their associated model parameters $\boldsymbol{\omega}_K$ (including π_k, μ_k, and σ_k, k = 1, 2, …, K). For example, Figure 1 shows an arbitrary PDF (see solid red line) that can be represented as a linear combination of three component Gaussian PDFs (see lines with triangles, squares, and circles) according to Eq. (1).

Determining the GMM given by Eq. (1) needs the information of $\boldsymbol{\omega}_K$, i.e., π_k, μ_k, and σ_k, k = 1, 2, …, K. In general, the optimal combination of $\boldsymbol{\omega}_K$ of the K component Gaussian PDFs for a given X data can be identified through Expectation-Maximization algorithm (e.g., Bishop 2006). However, such a GMM training procedure can easily lead to overfitting of the PDF of X data by increasing K, particularly as only a limited number of geotechnical data are available. Determining a proper value K is pivotal to training GMM. The next section develops a Bayesian learning method for training GMM under a probabilistic framework.

3 Bayesian Learning of Gaussian Mixture Model

3.1 Bayesian model class selection

GMMs with different numbers, K, of component Gaussian PDFs constitute a pool of model classes, each of which has a fixed value of K and is given by Eq. (1). Consider, for example, N_M candidate model classes with K varying from 1 to N_M. For the K-th model class M_K, the conditional probability of the corresponding GMM model with K component Gaussian PDFs given a set of test data, $\mathbf{D}$, is written as (e.g., Yuen 2010; Cao and Wang 2013, 2014):

$$P(\mathrm{M}_K \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mathrm{M}_K) P(\mathrm{M}_K)}{p(\mathbf{D})}, \quad K = 1, 2, \cdots, \mathrm{N_M} \tag{3}$$

$$p(\mathbf{D}) = \sum_{K=1}^{\mathrm{N_M}} p(\mathbf{D} \mid \mathrm{M}_K) P(\mathrm{M}_K) \tag{4}$$

where $p(\mathbf{D})$ is a normalizing constant that is independent of $\mathbf{D}$; $P(\mathrm{M}_K)$ is the prior probability of the model class M_K; $p(\mathbf{D}|\mathrm{M}_K)$ is the PDF of $\mathbf{D}$ given M_K, and it is referred to as "model evidence" under Bayesian model class selection framework. As there is no prevailing knowledge on model classes in the absence of data, the prior probabilities of all the candidate model classes are considered as equal, i.e., $P(\mathrm{M}_K)=1/\mathrm{N_M}$. As a result, $P(\mathrm{M}_K|\mathbf{D})$ is proportional to $p(\mathbf{D}|\mathrm{M}_K)$. Then, selecting the model class with the maximum value of $P(\mathrm{M}_K|\mathbf{D})$, i.e., the most probable model class, is equivalent to selecting the model class with the maximum value of $p(\mathbf{D}|\mathrm{M}_K)$.

According to the Theorem of Total Probability, the model evidence of M_K is written as (e.g., Yuen, 2010):

$$p(\mathbf{D} \mid \mathrm{M}_K) = \int_{\boldsymbol{\omega}_K} p(\mathbf{D} \mid \boldsymbol{\omega}_K, \mathrm{M}_K)\, p(\boldsymbol{\omega}_K \mid \mathrm{M}_K)\, \mathrm{d}\boldsymbol{\omega}_K \tag{5}$$

where $\boldsymbol{\omega}_K$ represents model parameters of M_K, and it is comprised of $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$, and $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_K\}$ of M_K, i.e., $\boldsymbol{\omega}_K = [\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}]$; $p(\mathbf{D}|\boldsymbol{\omega}_K,\mathrm{M}_K)$ is the likelihood function; $p(\boldsymbol{\omega}_K|\mathrm{M}_K)$ is the prior distribution of model parameters.

The likelihood function $p(\mathbf{D}|\boldsymbol{\omega}_K,\mathrm{M}_K)$ represents the PDF of $\mathbf{D}$ for a given M_K and its associated model parameters $\boldsymbol{\omega}_K$. Assuming that geotechnical data are statistically independent, the likelihood function is then calculated as:

$$P(\mathbf{D} \mid \boldsymbol{\omega}_K, \mathrm{M}_K) = \prod_{n=1}^{\mathrm{N_D}} p(X_n \mid \boldsymbol{\omega}_K, \mathrm{M}_K) \tag{6}$$

where $\mathbf{D} = \{X_n,\ n = 1, 2, \ldots, \mathrm{N_D}\}$ are a number of test data of X; and $p(X_n \mid \boldsymbol{\omega}_K, \mathrm{M}_K)$ is the conditional PDF value of X_n given $\boldsymbol{\omega}_K$ and M_K, and it is given by Eq. (1).

The prior distribution $p(\boldsymbol{\omega}_K|\mathrm{M}_K)$ reflects prior knowledge, such as engineering judgment and experience (e.g., Cao et al. 2016). For the case without prevailing prior knowledge, the prior distribution can be taken as a diffused one to represent the relatively non-informative prior knowledge (dos Santos Silva and Lopes 2008). Consider, for example, a uniform distribution as the prior for model parameter $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Assume that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are mutually independent in the absence of data. Then, the prior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ is expressed as:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma} \mid \mathrm{M}_K) = \begin{cases} \prod_{k=1}^{K} \dfrac{1}{\mu_{k,\max} - \mu_{k,\min}} \times \dfrac{1}{\sigma_{k,\max} - \sigma_{k,\min}} & \text{for } \mu_k \in (\mu_{k,\min}, \mu_{k,\max}), \sigma_k \in (\sigma_{k,\min}, \sigma_{k,\max}) \forall k \le K \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\mu_{k,\min}$ and $\mu_{k,\max}$ are minimum and maximum values of μ_k; $\sigma_{k,\min}$ and $\sigma_{k,\max}$ are minimum and maximum values of σ_k. The mixing coefficients $\boldsymbol{\pi}$ are considered following flat Dirichlet distribution, and it is expressed as (e.g., Bishop 2006; Cao et al. 2018):

$$p(\boldsymbol{\pi} \mid \mathrm{M}_K) = \Gamma(K) \tag{8}$$

where $\Gamma(\cdot)$ is the Gamma function evaluated at the value of K. Substituting Eqs. (4)-(8) into Eq. (3) gives the probability of M_K given a set of $\mathbf{D}$.

3.2 Bayesian model parameter identification

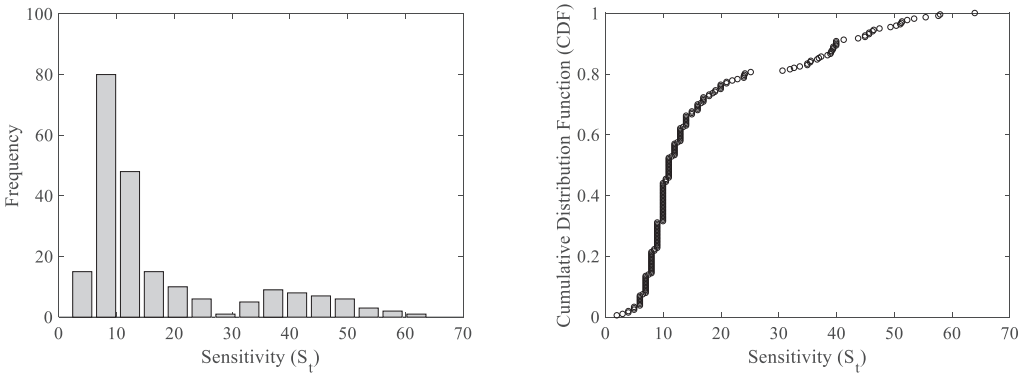
According to Bayes' Theorem, the posterior distribution $p(\boldsymbol{\omega}_K \mid \mathbf{D}, \mathrm{M}_K)$ of model parameters of M_K is expressed as (e.g., Yuen 2010; Cao and Wang 2013, 2014)

$$p(\boldsymbol{\omega}_K \mid \mathbf{D}, \mathrm{M}_K) = \frac{p(\mathbf{D} \mid \boldsymbol{\omega}_K, \mathrm{M}_K)\, p(\boldsymbol{\omega}_K \mid \mathrm{M}_K)}{p(\mathbf{D} \mid \mathrm{M}_K)} \tag{9}$$

where $p(\mathbf{D}|\mathrm{M}_K)$, $p(\mathbf{D}|\boldsymbol{\omega}_K,\mathrm{M}_K)$, and $p(\boldsymbol{\omega}_K|\mathrm{M}_K)$ are given by Eqs. (5)- (7), respectively. The dimension of posterior distribution of $\boldsymbol{\omega}_K$ given by Eq. (9) is equal to $3K$, and increases from 3 to $3\mathrm{N_M}$ as K increases from 1 to $\mathrm{N_M}$. Solving the posterior distribution of $\boldsymbol{\omega}_K$ might not be a trivial task because it involves a high-dimensional (e.g.,

Table 1. Ranges of model parameters used in prior distribution.

Parameters	Mean value of component Gaussian μ_k	Standard deviation of component Gaussian σ_k	Component weighting coefficient π_k
Possible Ranges	(0,100)	(0,30)	(0,1)



(a) Histogram. (b) Empirical cumulative distribution function.

Figure 2. Histogram and empirical cumulative distribution function of the sensitivity (S_t) of Finnish soft clays.

3K) integral in the normalizing constant $p(\mathbf{D}|\mathrm{M}_K)$, which is also the evidence of M_K in Eq. (3) for comparing model classes with different numbers of component Gaussian PDFs to represent the PDF of X in Eq. (1).

In this study, Bayesian updating with structural reliability methods (BUS) using Subset Simulation (SuS) (Straub and Papaioannou, 2015) is used to calculate the model evidence of M_K and to generate its corresponding posterior samples of $\boldsymbol{\omega}_K$ from Eq. (9). For the sake of conciseness, detailed algorithms and implementation procedures of applying BUS with SuS in model class selection and system identification problems are referred to Straub and Papaioannou (2015) and Cao et al. (2018). For implementing the BUS with SuS algorithm developed by Straub and Papaioannou (2015), the maximum value of the likelihood function $p(\mathbf{D}|\boldsymbol{\omega}_K,\mathrm{M}_K)$ for M_K is needed, which is pre-determined through the Expectation-Maximization algorithm (e.g., Bishop, 2006) in this study.

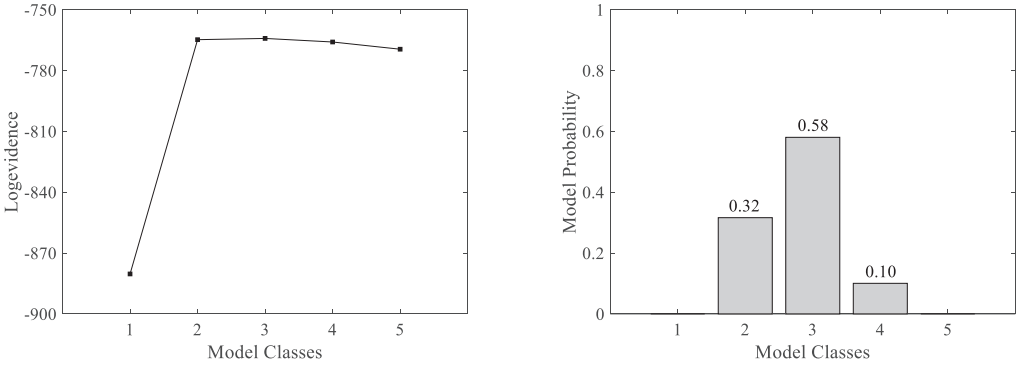
4 Illustrative Example

For illustration, the proposed approach is applied to geotechnical data in 304dB, which is compiled on a mission of developing publicly-accessible database for scientific research about inherent spatial variability, soil and rock properties, and risk assessment and management by ISSMGE TC304. In this study, a database, F-CLAY/7/216, that was developed for Finland soft sensitive clays by D'Ignazio et al. (2016) are used. The soft sensitive clays are widespread in the coastal areas of Finland. The significant variability of compressibility and sensitivity of Finnish soft clays renders geotechnical design a challenging task in Finland. The F-CLAY/7/216 database consists of 216 data points from different sites in Finland, each of which contains information of seven geotechnical parameters collected from comparable depths and site locations (D'Ignazio et al. 2016). In this section, the sensitivity (S_t) data of Finnish soft clays are used to illustrate the proposed method. Figure 2 shows the histogram and empirical cumulative distribution function (CDF) of S_t estimated from 216 S_t values provided by F-CLAY/7/216. It shows obviously that the probability distribution of S_t cannot be represented by a model with a single mode. Using 216 values of S_t as input of the proposed approach, a Gaussian mixture model of S_t is learned in this study.

Consider, for example, five candidate model classes (i.e., the maximum number of components of mixture model, $\mathrm{N_M} = 5$). The possible component number, K, of GMM is 1, 2, 3, 4, and 5. For all the model classes, the prior distributions of model parameters are provided by Eq. (7) and Eq. (8), where the possible ranges of model parameters are shown in Table 1. For populating the posterior distribution using BUS with SuS, 200,000 samples are simulated in each level of SuS and the conditional probability value, P_0, is taken as 0.1. Then, the proposed Bayesian learning method is used to determine the most probable number of Gaussian components and posterior distributions of their model parameters in this example.

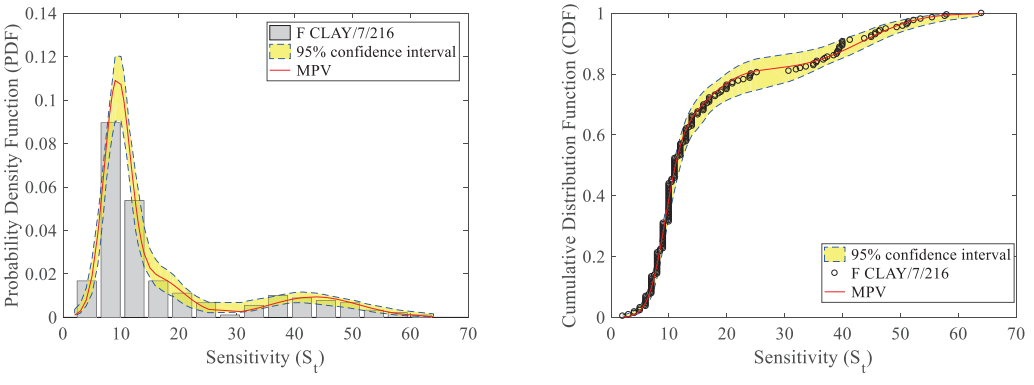
Table 2. Most probable values of model parameters of GMM with three components.

Component number, K	Most probable model parameters		
	μ_k	σ_k	π_k
1	15.71	5.00	0.23
2	9.29	2.27	0.59
3	43.45	7.77	0.18



(a) Logarithm of model evidence. (b) Model probability.

Figure 3. Logarithm of model evidence and model probability of candidate model classes



(a) Probability density function. (b) Cumulative distribution function.

Figure 4. Confidence interval and MPV of probability density function and cumulative distribution function of S_t.

Figures 3(a) and 3(b) show the logarithm of evidence and the occurrence probability of the five candidate model classes. The value of $\ln[p(\mathbf{D}|M_K)]$ increases from -880.30 to -764.17 as K increases from 1 to 3, and it then decreases from -763.16 to -769.49 as K further increases from 3 to 5. The model class M_3 with three components has the maximum value of $p(\mathbf{D}|M_K)$ of around 0.58 among all five model classes. Moreover, the model probabilities of M_2 and M_4 are around 0.32 and 0.10, respectively, and the occurrence probabilities of M_1 and M_5 are much smaller than those of M_2-M_4 and can be practically ignored. Therefore, the most probable value (MPV) of K among the five possible values is 3 in this example, and its corresponding MPVs of model parameters are summarized in Table 2, which maximizes the posterior distribution $p(\boldsymbol{\omega}_3|\mathbf{D}, M_3)$ of M_3.

Figures 4 (a) and 4(b) show the MPV of PDF and CDF determined by the MPVs of model parameters of M_3 by solid red line. It is found that the GMM PDF and CDF of S_t learned from the proposed approach fit reasonably well with the histogram and empirical CDF of S_t, respectively. Moreover, the proposed approach also provides posterior samples of model parameters of M_3, based on which 95% confidence interval of the GMM

PDF and CDF of S_t are obtained, as shown by the yellow zone between dashed lines. It is interesting to note that the 95% confidence interval of CDF learned from the proposed approach is relatively wide around $S_t = 30$, reflecting the fact of a lack of data points around the value. The proposed approach not only provides the most probable GMM of S_t but also is able to quantify its associated identification uncertainty.

5 Summary and Conclusions

This paper developed a Bayesian learning method that combines Bayesian model class selection with Gaussian mixture model (GMM) to identify a probabilistic model of geotechnical data. The proposed approach makes uses of GMM to represent arbitrary probability density function (PDF) of geotechnical parameters and determines the number of component Gaussian PDFs through Bayesian model class selection to avoid overfitting. Bayesian updating with structural reliability methods (BUS) using Subset Simulation (SuS) is applied to solving posterior distribution of each candidate model class and to calculate the model evidence for model classes selection. The proposed approach is illustrated through sensitivity (S_t) data of Finnish soft clays in 304Db, ISSMGE TC304 database (http://140.112.12.21/issmge/tc304.htm). Results showed that the proposed approach not only identifies the most probable model of S_t, but also quantifies its associated identification uncertainty, reflecting the degrees-of-belief in the GMM identified from the proposed approach. Although the proposed approach is developed and illustrated through univariate geotechnical data in this study, it can be extended to analyze multivariate geotechnical data, such as those compiled in 304dB, which is warranted in future study.

Acknowledgments

This work was supported by the National Key R&D Program of China (Project No. 2016YFC0800200), the National Natural Science Foundation of China (Project Nos. 51679174, 51779189, and 51879205), and Young Elite Scientists Sponsorship Program by CAST (Project No. 2017QNRC001). The financial support is gratefully acknowledged.

References

Baecher, G. B. and Christian, J. T. (2005). *Reliability and Statistics in Geotechnical Engineering*, John Wiley and Sons.

Bishop, C. M., (2006). Pattern recognition and machine learning. *Springer Science and Business Media*, LLC, USA.

Cao, Z. J. and Wang, Y. (2014). Bayesian model comparison and characterization of undrained shear strength. *Journal of Geotechnical and Geoenvironmental Engineering*, 140(6), 04014018.

Cao, Z. J. and Wang, Y. (2013). Bayesian approach for probabilistic site characterization using cone penetration tests. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(2), 267-276.

Cao, Z. J. and Wang, Y., and Li, D. Q. (2016). Quantification of prior knowledge in geotechnical site characterization. *Engineering Geology*, 203, 107-116.

Cao, Z. J., Zheng, S., Li, D. Q., and Phoon, K. K. (2019). Bayesian identification of soil stratigraphy based on soil behaviour type index. *Canadian Geotechnical Journal*, 56(4), 570-586.

D'Ignazio, M., Phoon, K. K., Tan, S. A., and Länsivaara, T. T. (2016). Correlations for undrained shear strength of Finnish soft clays. *Canadian Geotechnical Journal*, 53(10), 1628-1645.

dos Santos Silva, R., and Lopes, H. F. (2008). Copula, marginal distributions and model selection: A Bayesian note. *Statistics and Computing*, 18(3), 313-320.

Lumb, P. (1966). The variability of natural soils. *Canadian Geotechnical Journal*, 3(2), 74-97.

Lacasse, S. and Nadim, F.. (1997). Uncertainties in characterising soil properties. *Publikasjon - Norges Geotekniske Institutt*, 201, 49-75.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, Willey Series in Probability and Statistics.

Phoon, K. K. and Kulhawy, F. H. (1999). Characterization of geotechnical variability. *Canadian Geotechnical Journal*, 36(4), 612-624.

Phoon, K. K. and Kulhawy, F. H. (1999). Evaluation of geotechnical property variability. *Canadian Geotechnical Journal*, 36(4), 625-639.

Phoon, K. K. (2004). Towards reliability-based design for geotechnical engineering. *Special Lecture for Korean Geotechnical Society*, Seoul, 9, 1-23.

Straub, D. and Papaioannou, I. (2015). Bayesian updating with structural reliability methods. *Journal of Engineering Mechanics*, 141(3), 04014134.

Wang, Y., Zhao, T., and Cao, Z. J. (2015). Site-specific probability distribution of geotechnical properties. *Computers and Geotechnics*, 70, 159-168.

Wang, Y. and Cao, Z. J. (2013). Probabilistic characterization of Young's modulus of soil using equivalent samples. *Engineering Geology*, 159, 106-118.

Yuen, K. V. (2010). Recent developments of Bayesian model class selection and applications in civil engineering, *Structural Safety*, 32(5), 338-346.

ISSMGE TC304 database (http://140.112.12.21/issmge/tc304.htm)