Computers and Geotechnics 87 (2017) 229–240

Contents lists available at ScienceDirect

Computers and Geotechnics

journal homepage: www.elsevier.com/locate/compgeo

Research Paper

Impact of sample size on geotechnical probabilistic model identification

Xiao-Song Tang [a], Dian-Qing Li [a,*], Zi-Jun Cao [a], Kok-Kwang Phoon [b]

[a] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, 8 Donghu South Road, Wuhan 430072, PR China
[b] Department of Civil and Environmental Engineering, National University of Singapore, Blk E1A, #07-03, 1 Engineering Drive 2, Singapore 117576, Singapore

ARTICLE INFO

Article history:
Received 18 September 2016
Received in revised form 9 February 2017
Accepted 26 February 2017

Keywords:
Probabilistic models
Copulas
Statistical uncertainty
Sample size
Model identification

ABSTRACT

This paper aims to investigate the impact of sample size on geotechnical probabilistic model identification. First, the copula approach is presented to model the bivariate distribution of geotechnical parameters. Thereafter, the AIC scores are adopted to identify the best-fit marginal distribution and copula. Second, the variation of AIC scores because of small sample size is investigated using simulated data. Finally, the impact of the variation of AIC scores on identification of the best-fit marginal distribution and copula is examined. The minimum sample sizes for geotechnical data are also suggested to obtain a correct identification of the probabilistic models. The results indicate that the AIC scores estimated from a small sample exhibit large variation. The variation of the AIC scores has a significant impact on probabilistic model identification. The marginal distributions and copulas have a low percentage of correct identification when sample size is small. The percentages of correct identification for the marginal distributions and copulas increase with increasing sample size. The correlation coefficient between geotechnical parameters has a much larger impact on probabilistic model identification than the COV of geotechnical parameters. The suggested minimum sample sizes for geotechnical data are useful for guiding practical geotechnical site investigation.

1. Introduction

It is well known that the probabilistic models (i.e., marginal distribution for single parameter or joint probability distribution for multiple correlated parameters) for geotechnical parameters are essential inputs for geotechnical reliability analysis and risk assessments [27,3,8,22]. It is also widely accepted that there exist many examples for correlated geotechnical parameters in the literature (e.g., [14,26,9,6,5,34]). For example, the cohesion and friction angle of soils and rocks are commonly assumed negatively correlated [14,15,28–30,33,36]. The two curve-fitting parameters in a load-settlement curve of piles [9,16,17,12], and curve-fitting parameters in a soil-water characteristic curve [26] also have a strongly negative correlation. Furthermore, multiple soil parameters can be correlated with each other [4,6,5]. To achieve a realistic evaluation of geotechnical reliability and risk, the joint probability distribution of these parameters should be constructed.

Recently, the copula approach (e.g., [25]) provides a general and flexible way for modeling the joint probability distribution of correlated geotechnical parameters (e.g., [16–18,29,30,33,12,36,35]). In probability and statistics, a copula refers to a function that links a joint probability distribution to its one-dimensional marginal distributions. There are many copulas in the literature to characterize the dependence structure among variables such as Gaussian, *t*, Plackett, Frank, Clayton and Gumbel copulas (e.g., [25]). It is also clear that there is a variety of marginal distributions to describe the probabilistic properties for single variables such as normal, lognormal, Gumbel, Weibull and beta distributions (e.g., [2]). The copula approach constructs the joint probability distribution of geotechnical parameters by combing their marginal distributions with a copula function. In this study, probabilistic models refer to the marginal distributions and copula used to construct a joint probability distribution. Note that probabilistic models are uniquely characterized by their parameters and types. There are many methods available in the literature for estimating parameters in probabilistic models such as the method of moments, the maximum likelihood estimation (MLE) [2], and the Bayesian approach [31,32]. Different from the method of moments and MLE using geotechnical data only, the Bayesian approach uses both geotechnical data and prior information such as engineering judgement, local experience and published studies and reports to produce an estimation of the parameters [31,32]. Furthermore, the best-fit type of probabilistic models are usually identified using AIC scores [1] and goodness of fit (GOF) tests such as the Kolmogorov-Smimov (K-S) test, Anderson-Darling (A-D) test and

[*] Corresponding author.
E-mail address: dianqing@whu.edu.cn (D.-Q. Li).

http://dx.doi.org/10.1016/j.compgeo.2017.02.019
0266-352X/© 2017 Elsevier Ltd. All rights reserved.

Chi-square (χ^2) test from a pool of candidate probabilistic models (e.g., [21]).

According to statistics, the derived probabilistic models using the aforementioned approaches are accurate only when the sample size of geotechnical data is infinitely large. In geotechnical practice, geotechnical data are often of small sample size. The sample size of geotechnical data in a specific site is typically less than 30 for common geotechnical parameters [27,31,32,7]. The sample statistics estimated from a small sample exhibit large variation, which will induce uncertainty in the derived probabilistic models. In other words, the identified marginal distributions and copula based on a small sample may be incorrect [19,20]. In the literature, there are few studies that focus on characterizing the uncertainty in probabilistic models. Recently, Li et al. [19,20] proposed a bootstrap method to model the variation of the AIC scores and characterize the uncertainty in probabilistic models for geotechnical reliability analysis. However, the studies by Li et al. [19,20], only focused on a specified sample size. The impact of various sample sizes on probabilistic model identification has not been investigated. Furthermore, the question that what sample size is sufficient to obtain a correct identification of probabilistic models has not been answered.

This paper aims to investigate the impact of sample size on geotechnical probabilistic model identification. To achieve this goal, this article is organized as follows. In Section 2, the copula approach is presented to model the bivariate distribution of correlated geotechnical parameters. Thereafter, the AIC is presented to identify the best-fit marginal distribution and copula. In Section 3, the variation of the AIC scores for various marginal distributions and its impact on marginal distribution identification is investigated. The minimum sample sizes are suggested to obtain a correct identification of the best-fit marginal distribution. The variation of the AIC scores for various copulas and its impact on copula identification is presented in Section 4. The minimum sample sizes to obtain a correct identification of the best-fit copula are also suggested. Some discussions are provided in Section 5.

2. The copula approach for modeling a bivariate distribution

2.1. The copula approach

Let $F(x_1, x_2)$ be the joint cumulative distribution function (CDF) of two geotechnical parameters, X_1 and X_2. The marginal CDFs of X_1 and X_2 are denoted as $F_1(x_1)$ and $F_2(x_2)$, respectively. According to Sklar's theorem (e.g., [25]), $F(x_1, x_2)$ can be expressed in the following general form:

$$F(x_1,x_2) = C(F_1(x_1),F_2(x_2);\theta) = C(u_1,u_2;\theta) \tag{1}$$

where $C(u_1, u_2; \theta)$ is a bivariate copula function, and θ is a copula parameter describing the dependency between X_1 and X_2. As shown in Eq. (1), $F_1(x_1)$ and $F_2(x_2)$ are usually denoted as u_1 and u_2 ranging from 0 to 1. Therefore, both u_1 and u_2 are standard uniform variables, and $C(u_1, u_2; \theta)$ is essentially a bivariate probability distribution on $[0, 1]^2$ with uniform marginal probability distributions on [0, 1]. By taking derivatives of Eq. (1), the joint probability density function (PDF) of X_1 and X_2, $f(x_1, x_2)$, can be obtained as:

$$\begin{aligned} f(x_1,x_2) &= \frac{\partial^2 C(F_1(x_1),F_2(x_2);\theta)}{\partial F_1(x_1)\partial F_2(x_2)} \frac{\partial F_1(x_1)}{\partial x_1} \frac{\partial F_2(x_2)}{\partial x_2} \\ &= c(F_1(x_1),F_2(x_2);\theta) f_1(x_1) f_2(x_2) \\ &= c(u_1,u_2;\theta) f_1(x_1) f_2(x_2) \end{aligned} \tag{2}$$

where $f_1(x_1)$ and $f_2(x_2)$ are the marginal PDFs of X_1 and X_2, respectively; $c(u_1, u_2; \theta)$ is the bivariate copula density function associated with the bivariate copula function $C(u_1, u_2; \theta)$, which is given by

$$c(u_1,u_2;\theta) = \partial^2 C(u_1,u_2;\theta)/\partial u_1 \partial u_2 \tag{3}$$

Sklar's theorem states that a joint probability distribution can be expressed in terms of a copula function and its marginal distributions. Given the marginal distributions of X_1 and X_2, and the copula function describing the dependence structure between X_1 and X_2, the joint CDF and PDF of X_1 and X_2 can be obtained by using Eqs. (1) and (2). For example, assuming both X_1 and X_2 are normally distributed, their marginal CDFs, $F_1(x_1)$ and $F_2(x_2)$, can be respectively written as [2]:

$$F_1(x_1) = u_1 = \Phi\left(\frac{x_1 - \mu_1}{\sigma_1}\right) \tag{4}$$

and

$$F_2(x_2) = u_2 = \Phi\left(\frac{x_2 - \mu_2}{\sigma_2}\right) \tag{5}$$

where $\Phi(\cdot)$ is the univariate standard normal distribution function; μ_1 and σ_1 are the mean and standard deviation (SD) of X_1; μ_2 and σ_2 are the mean and SD of X_2. Similarly, assuming the dependence structure between X_1 and X_2 can be characterized by a Frank copula, its copula function, $C(u_1, u_2; \theta)$, has the following functional form:

$$C(u_1,u_2;\theta) = -\frac{1}{\theta} \ln\left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right] \tag{6}$$

Then, substituting Eqs. (4)–(6) into Eq. (1), the joint CDF, $F(x_1, x_2)$, of X_1 and X_2 can be obtained as:

$$\begin{aligned} F(x_1,x_2) = &-\frac{1}{\theta} \\ &\times \ln\left[1 + \left(e^{-\theta\Phi\left(\frac{x_1-\mu_1}{\sigma_1}\right)} - 1\right)\left(e^{-\theta\Phi\left(\frac{x_2-\mu_2}{\sigma_2}\right)} - 1\right)\Big/(e^{-\theta} - 1)\right] \end{aligned} \tag{7}$$

Therefore, modeling the joint probability distribution of X_1 and X_2 using the copula approach includes the following decoupled tasks: (1) determining the marginal distributions of X_1 and X_2, and (2) selecting a copula to describe the dependence structure between X_1 and X_2. The above two tasks using the data of X_1 and X_2 are detailed below.

2.2. Identification of the best-fit marginal distribution using AIC score

In geotechnical practice, the normal distribution truncated below zero (referred to as TruncNormal hereafter), lognormal distribution, and beta distribution are commonly adopted to fit the marginal distributions of geotechnical parameters (e.g., [33,21,19,20]). In this study, the above three distributions are selected as the set of candidate distributions to fit the marginal distributions of X_1 and X_2. These three distributions can ensure that the simulated data of X_1 and X_2 are positive, satisfying the requirements of positive geotechnical parameters [27]. Table 1 gives the PDFs, $f(x; p, q)$, for the three distributions, where (p, q) is a pair of distribution parameters. Note that parameters, a and b, in the PDF of the beta distribution are the lower bound and upper bound, and are set as zero and infinity in this study, respectively. The choice of zero and infinity as the lower bound and upper bound of the beta distribution is to model an arbitrary positive parameter. In general, the bounds should be chosen according to the physical bounds of the modeled parameter. For example, if the modeled parameter is related to the friction angle of soils and rocks, then the maximum value (upper bound) should be 90° instead of infinity.

Note that the mean and SD of a distribution will change when truncation is performed because the corresponding density also changes. In mathematics, there are two ways to express a trun-