Structural Safety 63 (2016) 21–32

Contents lists available at ScienceDirect

Structural Safety

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

Robust estimation of correlation coefficients among soil parameters under the multivariate normal framework

Jianye Ching [a,*], Kok-Kwang Phoon [b], Dian-Qing Li [c]

[a] Dept of Civil Engineering, National Taiwan University, Taiwan, ROC
[b] Dept of Civil and Environmental Engineering, National University of Singapore, Singapore
[c] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, PR China

ARTICLE INFO

Article history:
Received 4 March 2016
Received in revised form 12 July 2016
Accepted 12 July 2016

Keywords:
Soil properties
Correlation
Multivariate normal distribution
Statistical uncertainty

ABSTRACT

Based on limited amount of multivariate soil data $\underline{Y}$, it is only possible to reliably estimate the marginal distributions and the correlations. A common practical approach of constructing the multivariate probability distribution of $\underline{Y}$ is to transform $\underline{Y}$ into standard normal data $\underline{X}$ and construct the multivariate standard normal distribution for $\underline{X}$. This method is called the translation method. Its success depends on whether the Pearson product-moment correlations (δ_{ij}) for $\underline{X}$ can be robustly estimated. This paper investigates the robustness for four methods of estimating δ_{ij}. The emphasis is on the statistical uncertainty in the estimated δ_{ij} when the amount of soil data is limited. It is found that the well known method that maps the Pearson correlations for $\underline{Y}$ to δ_{ij} is the least robust, suffering the most significant statistical uncertainty. The causes for this non-robustness are investigated. The two methods that map the Spearman and Kendall rank correlations for $\underline{Y}$ to δ_{ij} are quite robust. The method that converts $\underline{Y}$ to $\underline{X}$ and directly estimates δ_{ij} is also robust as long as the conversion is based on properly chosen marginal distributions.

1. Introduction

Geotechnical data are multivariate in their nature. For instance, when borehole samples are drawn, SPT-N values are usually available; moreover, the information regarding unit weight, plasticity index (PI), liquid limit (LL) and water content can be quickly obtained through laboratory tests. Many of these test indices may be simultaneously correlated to a design soil parameter such as the undrained shear strength (s_u). Some multivariate soil databases have been compiled in recent studies [2,4,6,8]and multivariate probability distribution models have been constructed. Table 1 shows these databases, labeled as (soil type)/(number of parameters of interest)/(number of data points). With the multivariate distribution models, these studies showed that it is possible to reduce the uncertainty in the design soil parameter by incorporating multiple site investigation information. This reduction in uncertainty can further translate to actual savings in design dimensions under the reliability-based design framework [9]. This more explicit link between site investigation efforts and possible design savings is a distinctive and important subject in geotechnical engineering [20,35,36].

In practice, it is not possible to construct the exact multivariate distribution based on limited amount of data. The available information is typically limited to the marginal distributions and the correlations only [30,24]. Given the marginal distributions and the correlations of a set of parameters of interest, the underlying multivariate distribution is not unique [24]. A common practical approach of constructing the multivariate distribution is to transform the non-normal data into standard normal data and construct the underlying multivariate normal distribution. To be specific, let $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)$ be the multivariate geotechnical parameters of interest. In general, Y_i is non-normal, and the following CDF transform can be adopted to transform Y_i into standard normal variable X_i:

$$X_i = \Phi^{-1}[F_i(Y_i)] \quad Y_i = F_i^{-1}[\Phi(X_i)] \tag{1}$$

where F_i is the cumulative density function (CDF) for Y_i; Φ is the standard normal CDF; Φ^{-1} is the inverse function for Φ; F_i^{-1} is the inverse function for F_i. Furthermore, $\underline{X} = (X_1, X_2, \ldots, X_n)$ is assumed to follow the multivariate normal distribution with the Pearson product-moment correlation δ_{ij} between (X_i, X_j). This method of constructing multivariate distribution has broad applications in the literature [29,19,30,1,24]. It is called the "Nataf model" [31] in Liu and Der Kiureghian [30], the "NORTA (NORmal To Anything)

* Corresponding author.
E-mail address: jyching@gmail.com (J. Ching).

http://dx.doi.org/10.1016/j.strusafe.2016.07.002
0167-4730/

Table 1
Soil databases.

Database	Reference	Parameters of interest	# sites/studies	Marginal PDF	Method of determining δ_{ij}
CLAY/5/345	Ching and Phoon [2]	LI, s_u, s_u^{re}, σ'_p, σ'_v	37 sites	Lognormal	Method XP
CLAY/6/535	Ching et al. [8]	s_u/σ'_v, OCR, $(q_t-\sigma_v)/\sigma'_v$, $(q_t-u_2)/\sigma'_v$, $(u_2-u_0)/\sigma'_v$, B_q	40 sites	Johnson	Method XP
CLAY/7/6310	Ching and Phoon [4]	s_u under 7 different test modes	164 studies	Lognormal	Method XP
CLAY/10/7490	Ching and Phoon [6]	LL, PI, LI, σ'_v/P_a, σ'_p/P_a, s_u/σ'_v, S_t, $(q_t-\sigma_v)/\sigma'_v$, $(q_t-u_2)/\sigma'_v$, B_q	251 studies	Johnson	Method XP

LL: liquid limit; PI: plasticity index; LI: liquidity index; s_u: undrained shear strength; s_u^{re}: remolded s_u; σ'_p: preconsolidation stress; σ'_v: vertical effective stress; σ_v: vertical total stress; OCR: overconsolidation ratio; q_t: corrected cone tip resistance; u_2: pore pressure behind the cone; u_0: static pore pressure; B_q: CPTU pore pressure parameter; P_a: one atmosphere pressure; S_t: sensitivity. Method XP refers to the method of estimating the Pearson moment-product correlation (δ_{ij}) by transforming the non-normal soil data into standard normal variables.

distribution" in Cario and Nelson [1], and the "translation method" in Johnson [22] and Li et al. [24]. Some special cases of such models include the multivariate lognormal model proposed by Johnson and Ramberg [21] and the bivariate Johnson model proposed by Johnson [22], which is later extended to the multivariate Johnson model by Stanfield et al. [38]. For bivariate geotechnical data, the copula theory has been widely used for constructing bivariate distributions [25,26,28,27,39,40,41,18,17]. With the copula theory, it is possible to go beyond the bivariate normal distribution framework. However, there are only limited studies applying copulas to multivariate distributions with dimension more than 2 ($n > 2$), because only the elliptical copulas (e.g., Gaussian copula and t copula) have practical n-dimensional generalizations. The current study focuses on the multivariate normal distribution framework. This multivariate normal distribution framework will be referred to as the "translation method" and the multivariate model will be referred to as the "translation model" in the following.

The success of the translation model depends on whether the marginal probability density functions (PDF) and the Pearson product-moment correlations δ_{ij} can be reliably estimated. The goodness-of-fit of the marginal PDFs can be assessed through classical statistical tests such as the Chi-squared and K-S tests [12]. The focus of the current paper is on the robust estimation of the Pearson correlation δ_{ij} between each (X_i, X_j) pair. The translation method is by no means a complete framework. Not every multivariate distribution can be represented as a translation model. Therefore, the translation model should be considered as an approximate albeit practical model. Li et al. [24] investigated the performance of the translation model in approximating a non-translation model. The main goal was to verify the effectiveness of the translation model when the target multivariate distribution model is beyond the multivariate normal framework. Nonetheless, the purpose of the current paper is different. The purpose is to investigate the *statistical uncertainty* in the estimated δ_{ij}. Due to the limited amount of the geotechnical data, the estimated δ_{ij} is not identical to the actual δ_{ij}. The discrepancy is the statistical uncertainty. In this study, four methods of estimating δ_{ij} will be investigated. The one with the least statistical uncertainty should be considered as the most robust method. Note that these four methods will produce the same correlation coefficients if the data are produced by a translation model and the sample size is infinitely large.

To conduct this investigation, a translation model that is *within* the multivariate normal framework with marginal PDFs and correlation matrix is adopted to simulate $\underline{Y}$ data. This translation model was constructed by Ching and Phoon [2] based on the database CLAY/5/345 in Table 1. The simulated $\underline{Y}$ data are adopted to estimate δ_{ij} by the four methods. The discrepancy between the actual and estimated δ_{ij} can then be quantified, and conclusions regarding the robustness of each method will be given. To confirm the applicability of the conclusions with respect to real soil databases, further comparisons among the four methods will be conducted for the four real soil databases shown in Table 1.

2. Methods for estimating δ_{ij}

In the literature, there are at least four methods for estimating δ_{ij}. They are denoted by Method P, Method S, Method K, and Method XP below:

1. Method P. This method is the most common method adopted in the literature (e.g., [29,30,1,25]. It is based on the Pearson product-moment correlation between (Y_i, Y_j), denoted by ρ_{ij}, which can be estimated using the following equation:

$$\rho_{ij} \approx \frac{\frac{1}{N-1}\sum_{k=1}^{N}\left(Y_i^{(k)}-m_i\right)\cdot\left(Y_j^{(k)}-m_j\right)}{\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(Y_i^{(k)}-m_i\right)^2\times\frac{1}{N-1}\sum_{k=1}^{N}\left(Y_j^{(k)}-m_j\right)^2}} \tag{2}$$

where the superscript (k) is the sample index; m_i is the sample mean of Y_i; N is the total number of data points. To implement Method P, ρ_{ij} is first estimated from the soil data of (Y_i, Y_j), then δ_{ij} can be found by solving the following integral equation [29,30,25]:

$$\rho_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{F_i^{-1}[\Phi(x_i)]-\mu_i}{\sigma_i}\right)\left(\frac{F_j^{-1}[\Phi(x_j)]-\mu_j}{\sigma_j}\right)\times\frac{1}{2\pi\sqrt{1-\delta_{ij}^2}}\exp\left\{-\frac{x_i^2-2\delta_{ij}x_ix_j+x_j^2}{2(1-\delta_{ij}^2)}\right\}dx_idx_j \tag{3}$$

where μ_i and σ_i are the mean value and standard deviation for Y_i. If $(Y_1, Y_2, \ldots, Y_n)$ are multivariate lognormal, i.e., $Y_i = \exp(X_i)$, Eq. (3) has the following analytical form:

$$\delta_{ij} = \frac{\ln\left[1+\rho_{ij}\times(\sigma_i/\mu_i)\times(\sigma_j/\mu_j)\right]}{\sqrt{\ln\left[1+(\sigma_i/\mu_i)^2\right]}\times\sqrt{\ln\left[1+(\sigma_j/\mu_j)^2\right]}} \tag{4}$$

2. Method S. This method is adopted in Li et al. [24]. It is based on the Spearman rank correlation between (Y_i, Y_j), denoted by r_{ij}, which is defined to be the Pearson correlation between $[F_i(Y_i), F_j(Y_j)]$, where F_i is the CDF for Y_i. r_{ij} can be estimated as the Pearson correlation between the ranks of (Y_i, Y_j), namely using Eq. (2) but the (Y_i, Y_j) data are replaced by their ranks. Because $F_i(Y_i) = \Phi(X_i)$ (see Eq. (1)), it is clear that the Pearson correlation between $[F_i(Y_i), F_j(Y_j)]$ is identical to that between $[\Phi(X_i), \Phi(X_j)]$. This implies that the Spearman correlation between (Y_i, Y_j) is identical to that between (X_i, X_j). Moreover, *for bivariate normal* (X_i, X_j), their Pearson and Spearman correlations are related by the following equation [16]:

$$\delta_{ij} = 2\sin\left(\frac{\pi}{6}\times r_{ij}\right) \tag{5}$$

Eq. (5) does not apply to any bivariate distribution, although the error of doing so has not been studied due to the difficulty of simulating genuinely non-normal multivariate data, i.e., data that deviate significantly from those produced by the Nataf or