Bayesian identification of random field model using indirect test data

Mi Tian a, Dian-Qing Li a, Zi-Jun Cao a,⁎, Kok-Kwang Phoon b, Yu Wang c

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, BII, E1A, #07-03, 1 Engineering Drive 2, Singapore 117576, Singapore
c Department of Architecture and Civil Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

A R T I C L E   I N F O

Keywords:
Inherent spatial variability
Effective friction angle
Cone penetration test
Bayesian approaches
Markov Chain Monte Carlo Simulation
Gaussian copula

A B S T R A C T

Inherent spatial variability (ISV) of design soil properties (e.g., effective friction angle \( \phi' \)) can be incorporated into probability-based geotechnical analyses and designs using random field models. Defining a random field model includes determination of random field parameters (i.e., mean \( \mu \), standard deviation \( \sigma \), and scale of fluctuation \( \lambda \)) and the correlation function that specifies the spatial correlation of the concerned design soil property (e.g., \( \phi' \)) at different locations. This is, however, a challenging task at a given site due to a lack of direct test data of design soil properties and various uncertainties (e.g., transformation uncertainty) arising during site investigation. This paper develops Bayesian approaches for probabilistic characterization of the ISV of \( \phi' \) using indirect test data (i.e., cone penetration test (CPT) data) and prior knowledge, which identify random field parameters of \( \phi' \) through Markov Chain Monte Carlo Simulation (MCMCS) and, simultaneously, make use of Gaussian copula to select the most probable correlation function \( M^* \) among a pool of candidate correlation functions based on MCMCS samples. The proposed Bayesian approaches account, rationally and transparently, for the transformation uncertainty associated with the transformation model between \( \phi' \) and CPT data. The proposed approaches are illustrated and validated using real-life and simulated CPT data. Results show that the proposed approaches properly identify the random field model (including \( \mu, \sigma, \lambda, \) and \( M^* \)) of \( \phi' \) using project-specific CPT data, and the random field parameters of \( \phi' \) depend on the correlation function used to interpret CPT data. In addition, the suitability of MCMCS in Bayesian probabilistic characterization of soil properties is highlighted, particularly for the cases with a limited number of test data.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Inherent spatial variability (ISV) of soils is one of major sources of uncertainties in soil properties (e.g., Phoon and Kulhawy, 1999a; Baecher and Christian, 2003; Wang et al., 2016). It can be incorporated into probability-based geotechnical analyses and designs through random field theory (e.g., Fenton and Griffiths, 2008; Vanmarcke, 2010; Gong et al., 2014; Jamshidi Chenari and Alaie, 2015; Li et al., 2015a, 2015b, 2016a, 2016b). A random field model probabilistically characterizes the ISV through a set of random field parameters (i.e., mean \( \mu \), standard deviation \( \sigma \), and scale of fluctuation \( \lambda \)) and a correlation function (such as those shown in Fig. 1) (e.g., Fenton, 1999; Fenton and Griffiths, 2008; Lloret-Cabot et al., 2014; Kasama and Whittle, 2016). Determining the random field parameters and the correlation function of design soil properties, which are directly used in geotechnical designs (e.g., effective friction angle \( \phi' \)), at a site is, therefore, a necessary prerequisite for probabilistic characterization of ISV of soil properties at the site. This is, however, a challenging task in geotechnical practice.

Consider, for example, probabilistic characterization of the ISV of the effective friction angle \( \phi' \). Values of \( \phi' \) in a soil layer can be directly measured from laboratory tests (e.g., triaxial tests) on soil samples retrieved from boreholes in a discrete manner. The number of direct measurements of \( \phi' \) in a soil layer is usually too sparse to generate meaningful statistics and correlation function because a large number of laboratory tests are costly. On the other hand, \( \phi' \) can be indirectly estimated using fast and economical in-situ tests (e.g., cone penetration test (CPT)) through transformation models (e.g., the empirical regression between normalized cone tip resistance \( q \) from CPT and \( \phi' \), as shown in Fig. 2). The transformation model is not a perfect relationship but is associated with uncertainties/dispersion about a mean trend, namely “transformation uncertainty” (e.g., Phoon and Kulhawy, 1999b), which shall be rationally considered when using indirect test data (e.g., CPT data) to characterize the ISV of \( \phi' \). This can be formulated as an inverse analysis problem under a Bayesian framework (Wang et al., 2016).

Wang et al. (2010) and Cao and Wang (2013) proposed Bayesian approaches to inversely infer the random field model parameters (i.e., \( \mu, \sigma, \lambda \)) of \( \phi' \).
proposed a Bayesian approach to select a proper correlation function which is unknown prior to site investigation. Cao and Wang (2014a) approaches need to prescribe a correlation function before the analysis, certainty in an explicit and rational manner. However, these Bayesian (After Kulhawy and Mayne, 1990; Phoon and Kulhawy, 1999b; Wang et al., 2010).

Fig. 2. Four commonly-used correlation functions in geotechnical engineering (After Phoon et al., 2003).

and λ) of \( \psi' \) using CPT data, which account for the transformation uncertainty in an explicit and rational manner. However, these Bayesian approaches need to prescribe a correlation function before the analysis, which is unknown prior to site investigation. Cao and Wang (2014a) proposed a Bayesian approach to select a proper correlation function of \( q \) among a pool of candidate correlation functions, in which the random field parameters of \( q \) are treated as nuisance parameters and the transformation uncertainty is not involved. How to use a limited number of indirect test data (e.g., CPT data) to identify the random field parameters and simultaneously select an appropriate correlation function for probabilistic characterization of ISV of design soil properties (e.g., \( \psi' \)) at a specific site remains an outstanding challenge. In addition, effects of the correlation function on the random field parameters are also not clear.

This paper develops Bayesian approaches that identify the random field parameters (i.e., \( \mu, \sigma, \) and \( \lambda \)) of \( \psi' \) and simultaneously select the most probable correlation function of \( \psi' \) among a pool of candidates based on a limited number of project-specific CPT data and site information available prior to the project, namely “prior knowledge”. The information from CPT data and prior knowledge is systematically integrated as the posterior knowledge on \( \mu, \sigma, \) and \( \lambda \) under a Bayesian framework, which is quantitatively reflected by their posterior distribution. The computational complexity in solving the posterior distribution has been considered as one key limitation of Bayesian methods (e.g., Zhang et al., 2009). To bypass the computational complexity, the posterior distribution can be solved by Laplace asymptotic approximation method (LAAM) when it is well approximated by a Gaussian distribution (e.g., Wang et al., 2010; Cao and Wang, 2013; Ching et al., 2016). When there are only a limited number of test data, which is often the case in geotechnical engineering, the Gaussian approximation might not be valid. In such a case, Markov Chain Monte Carlo Simulation (MCMCS) provides a more appropriate tool to obtain the posterior knowledge in Bayesian analysis by generating random samples of model parameters concerned (e.g., \( \mu, \sigma, \) and \( \lambda \)) from the posterior distribution (e.g., Zhang et al., 2010, 2012; Wang and Cao, 2013, Jiang et al., 2013; Peng et al., 2014; Cao and Wang, 2014b; Kelly and Huang, 2015; Huang et al., 2016; Ching et al., 2016). Among various MCMCS algorithms, Metropolis–Hastings (M–H) algorithm (e.g., Metropolis et al., 1953; Hastings, 1970) is widely used for its simplicity. However, it has a key limitation that M-H algorithm does not give the likelihood of test data for a given model, which is often referred to as the “evidence” on the given model provided by test data in Bayesian model selection problems. This limitation makes the M–H algorithm infeasible in model selection problems (e.g., Bayesian selection of the most probable correlation function).

This paper removes the abovementioned limitations of M–H algorithm using Gaussian copula and selects the most probable correlation function among a pool of candidates (e.g., those shown in Fig. 1) based on MCMCS samples generated by M–H algorithm for candidate correlation functions. In addition, the proposed approaches also provide insights into effects of correlation functions on random field parameters. The paper starts with the development of the proposed Bayesian approaches, followed by a brief description of their implementation. Finally, the proposed approaches are illustrated and validated using real-life and simulated CPT data.

2. Bayesian identification of random field parameters

Random field theory (Vanmarcke, 2010) is used to explicitly model the ISV of \( \psi' \) within a statistically homogenous sand layer in this study, by which \( \psi' \) at different depths are modeled by a series of spatially correlated normal variables with a mean \( \mu \) and standard deviation \( \sigma \) (i.e., a one-dimensional stationary normal random field). The spatial correlation between variations of \( \psi' \) at different depths is then specified by the scale of fluctuation \( \lambda \) and a correlation function \( M \). Examples of correlation functions include the single exponential correlation function (SECF), binary noise correlation function (BNCF), second order Markov correlation function (SMCF), and squared exponential correlation function (SQEFC), as shown in Fig. 1. Note that the correlation function \( M \) is assumed to take a specific form (e.g., one of the correlation functions shown in Fig. 1) in this section, but its corresponding \( \lambda \) value is unknown herein. A proper form of the correlation function will be determined among a pool of candidates (e.g., those shown in Fig. 1) by a Bayesian model selection approach in Section 3 entitled “Bayesian selection of spatial correlation function using MCMCS samples”.

For a given correlation function \( M \), the stationary normal random field of \( \psi' \) is uniquely represented by the random field parameters \( \mathbf{X} \), i.e., \( [\mu, \sigma, \lambda] \). For a given set of prior knowledge and CPT data \( \xi \), there are various possible values of random field parameters, and their respective plausibility can be quantified by the posterior distribution \( P(\mathbf{X}|\xi, M) \) under a Bayesian framework, where the condition on \( M \) indicates that the correlation function is assumed to take a specific form.

![Fig. 1](image1.png)

**Fig. 1.** Four commonly-used correlation functions in geotechnical engineering (After Phoon et al., 2003).

![Fig. 2](image2.png)

**Fig. 2.** Regression between effective friction angle and normalized cone tip resistance (After Kulhawy and Mayne, 1990; Phoon and Kulhawy, 1999b; Wang et al., 2010).
Using Bayes’ Theorem (e.g., Ang and Tang, 2007; Yuen, 2010a, 2010b; Wang et al., 2016), \( P(\xi | X, M) \) is written as:

\[
P(\xi | X, M) = K^{-1}P(\xi | X, M)P(X | M)
\]

(1)

in which \( \xi \) is a set of values of the logarithm of \( q \) measured at \( n \) different depths \( D_1, D_2, \ldots, D_n \) in a sand layer by CPT, i.e., \( (\xi(D_1), \xi(D_2), \ldots, \xi(D_n))^T \). \( K \) is a normalizing constant that is independent of \( X \) for a given \( M \) and it is calculated as \( |P(\xi | X, M)P(X | M)dX| \). \( P(\xi | X, M) \) is the likelihood function reflecting the model fit with the CPT data for a given \( M \) and random field parameters \( X \), and its formulation is provided in the next subsection; and \( P(X | M) \) is the prior distribution of random field parameters that reflects the available prior knowledge on \( X \) in the absence of data. When there is no prevailing knowledge on \( X \) (e.g., only the possible ranges of \( X \) are available), a relatively non-informative prior distribution can be used to reflect the prior knowledge, e.g., a joint uniform distribution given by (e.g., Wang and Cao, 2013; Cao et al., 2016):

\[
P(X | M) = \begin{cases} 1 & \text{for } \mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}, \sigma_{\text{min}} \leq \sigma \leq \sigma_{\text{max}}, \lambda_{\text{min}} \leq \lambda \leq \lambda_{\text{max}} \\
0 & \text{otherwise} \end{cases}
\]

(2)

in which \( \mu_{\text{min}}, \sigma_{\text{min}}, \lambda_{\text{min}} \) are the minimum values of \( \mu, \sigma, \) and \( \lambda \), respectively; and \( \mu_{\text{max}}, \sigma_{\text{max}}, \lambda_{\text{max}} \) are the maximum values of \( \mu, \sigma, \) and \( \lambda \), respectively. In the case where local information (e.g., local experience and engineering judgement) is available at a site, the prior knowledge becomes more informative. Both prior information from the literature (e.g., possible ranges of \( X \)) and local site information shall be considered to determine the prior distribution of random field parameters, which might provide more sophisticated types of prior distributions (e.g., an arbitrary histogram type of prior distribution) (Cao et al., 2016). This may further complicate the posterior distribution in Eq. (1) and lead to difficulties in expressing it analytically and explicitly.

2.1. Likelihood function

Since the ISV of \( \psi \) along the depth in a sand layer is modeled by a one-dimensional stationary normal random field in this study, \( \psi(D) \) at \( n \) different depths can be represented by a normal random vector \( \psi(D) \) (i.e., \( \psi(D_1), \psi(D_2), \ldots, \psi(D_n))^T \), which is written as (e.g., Wang et al., 2010):

\[
\psi(D) = \mu + \Lambda^{1/2} Z
\]

(3)

in which \( \psi \) is a vector with all components equal to unit; \( Z \) is a standard normal vector with \( n \) independent components; and \( \Lambda \) is a \( n \)-by-\( n \) upper-triangular matrix obtained from Cholesky decomposition of the correlation matrix \( R \) of \( \psi(D) \). \( R \) is also a \( n \)-by-\( n \) positive definite matrix. Its \((i,j)\)-th entry is the correlation coefficient \( \rho(D_i, D_j) \) between \( \psi(D_i) \) and \( \psi(D_j) \) at depths \( D_i, D_j \) with a separate distance \( D_i - D_j \) that is calculated as \(|D_i - D_j|\). In the section, \( \rho(D_i, D_j) \) is calculated using a prescribed correlation function \( M \) such as SECF, BNF, SMCF, and SQCEF.

As shown in Fig. 2, \( \psi \) can be estimated from CPT data by a semi-log empirical regression model developed by Kulhawy and Mayne (1990). Using the regression model shown in Fig. 2 and Eq. (3), the CPT data \( \xi \) in a statistically homogenous sand layer is formulated as a normal random vector with a mean vector \((a + b)\) and a covariance matrix \( \Sigma \) equal to \( a^2 \sigma^2 R + \sigma^2 I \). Herein, \( a \) and \( b \) are taken as 0.209 and \(-3.684 \), respectively; \( \sigma \) reflects the transformation uncertainty associated with the regression model shown in Fig. 2, and it is equal to 0.586; and \( I \) is a \( n \)-by-\( n \) identity matrix. Details of the formulation and discussions on \( \xi \) are referred to Wang et al. (2010) and Cao and Wang (2013). Then, the probability density function (PDF) of \( \xi \) is given by a joint normal distribution:

\[
P\left(\xi | X, M\right) = \left(2\pi \right)^{-n/2} \left| \Sigma \right|^{-1/2} \exp \left\{-\frac{1}{2} (\xi - (a + b))^T \Sigma^{-1} (\xi - (a + b)) \right\}
\]

(4)

Eq. (4) is then taken as the likelihood function in Eq. (1). After the prior distribution (e.g., Eq. (2)) and likelihood function (e.g., Eq. (4)) are obtained, they are substituted into Eq. (1) to obtain the posterior distribution \( P(\xi | X, M) \). To avoid the computational complexity in solving the posterior distribution that is considered as one key limitation of Bayesian methods in literature (e.g., Zhang et al., 2009; Wang et al., 2010), the next subsection uses MCMCS to generate a large number of \( X \) samples to describe the posterior distribution numerically.

2.2. Markov Chain Monte Carlo Simulation

MCMCS is a numerical process that simulates a sequence of random samples of target random variable (e.g., \( X \)) as a Markov Chain with the PDF (\( P(\xi | X, M) \)) of the random variables as the Markov Chain’s limiting stationary distribution (e.g., Beck and Au, 2002; Robert and Casella, 2004; Wang and Cao, 2013). Among various MCMCS algorithms, M-H algorithm is one of the simplest methods and has several successful applications in geotechnical engineering (e.g., Chiu et al., 2012; Wang and Cao, 2013; Jiang et al., 2013; Cao and Wang, 2014b). In this study, M-H algorithm is used to generate a large number (i.e., \( N_{CF} \)) of \( X \) (i.e., \( \mu, \sigma, \lambda \)) samples from Eq. (1). The samples of \( X \) obtained after the Markov Chain reaches its stationary condition are taken as appropriate samples for numerically representing the posterior distribution, and they are subsequently used to construct the posterior PDFs and cumulative distribution functions (CDFs) of \( \mu, \sigma, \) and \( \lambda \) through conventional statistical analyses. More details of M-H algorithm are referred to Wang and Cao (2013); Jiang et al. (2013), and Au and Wang (2014).

One of the major advantages of using MCMCS to solve the Bayesian equation (e.g., Eq. (1)) is that the calculation of the normalizing constant \( K \) that involves a multi-dimensional integration is avoided (e.g., Beck and Au, 2002; Wang and Cao, 2013). However, \( K \) plays a key role in selecting an appropriate model (e.g., a spatial correlation function) from a pool of candidates under a Bayesian framework (e.g., Yuen, 2010a, 2010b; Cao and Wang, 2014a, 2014b; Wang and Aladjevar, 2015), as discussed in the next subsection.

3. Bayesian selection of spatial correlation function using MCMCS samples

3.1. Bayesian model selection

The spatial correlation function \( M \) is considered as a prescribed one with unknown \( \lambda \) in the previous section. This section develops a Bayesian model selection method to determine the most probable correlation function \( M^* \) among a pool of candidate correlation functions for a given set of \( \xi \). Consider a number (i.e., \( N_{CF} \)) of candidate correlation functions \( M_k (k = 1, 2, \ldots, N_{CF}) \), such as SECF, BNF, SMCF, and SQCEF shown in Fig. 1. For a given set of \( \xi \), the plausibility of the \( k \)-th candidate correlation function \( M_k \) is quantified by \( P(M_k | \xi) \). Using Bayes’ theorem, \( P(M_k | \xi) \) is given by (e.g., Yuen, 2010a, 2010b; Cao and Wang, 2014a, 2014b):

\[
P(M_k | \xi) = P(\xi | M_k) P(M_k) / P(\xi) \]

(5)

where \( P(\xi) \) is a normalizing constant that is independent of \( M_k \), and it is calculated as \( \sum_{k=1}^{N_{CF}} P(\xi | M_k) P(M_k) \). \( P(M_k) \) is the prior probability of \( M_k \); and \( P(\xi | M_k) \) is the conditional PDF of \( \xi \) for a given correlation function \( M_k \).
and it quantifies the information on $M_0$ provided by CPT data and is referred to as “evidence” for $M_0$ provided by $\hat{\xi}$.

In the case of no prevailing prior knowledge on $M_0$, the $N_{CF}$ candidate correlation functions are considered to have the same prior probability $P(M_0)$, i.e., $1/N_{CF}$. $P(M_0)$ is, therefore, a constant for $N_{CF}$ candidate correlation functions. Since both $P(M_0)$ and $P(\hat{\xi})$ in Eq. (5) are constants, the $P(M_0|\hat{\xi})$ is proportional to the evidence $P(\hat{\xi}|M_0)$. Because the most probable correlation function $M^*$ is defined as the one with the maximum value of $P(M_0|\hat{\xi})$, it also has the maximum value of $P(\hat{\xi}|M_0)$. In other words, $M^*$ can be determined by comparing the values of $P(\hat{\xi}|M_0)$ of the $N_{CF}$ candidate correlation functions.

Note that $P(\hat{\xi}|M_0)$ is actually the normalizing constant $K$ in Eq. (1) by setting $M$ as $M_0$, which involves a multi-dimensional integration. As discussed in the previous section, there is no need to calculate $K$ (i.e., $P(\hat{\xi}|M_0)$) when using MCMCS to solve the posterior distribution of $X$ in Eq. (1) for $M_0$. However, it is needed herein for comparing different correlation functions. The next subsection makes use of MCMCS samples of $X$ generated from Eq. (1) for $M_0$ and Gaussian copula to evaluate $P(\hat{\xi}$.

---

**Fig. 3.** Flow chart for the implementation of the proposed Bayesian approaches.
distribution in Bayesian analyses. In the context of copula theory, a distribution that uses copula to describe the multi-dimensional posterior distribution can be written as (e.g., Yuen, 2010b):

$$\text{ln}(P(\xi|X)) = \sum_{i=1}^{N_p} \text{ln}(P(\xi_i|X_i))$$

where $P(\xi|X)$ is the likelihood function that is calculated using Eq. (4) by setting $M$ as $M_2$ and $P(X|M_2)$ is the prior distribution of $X$ given by Eq. (2). Then, the logarithm of $P(\xi|X)$ can be written as (e.g., Yuen, 2010b):

$$\text{ln}(P(\xi|X)) = \sum_{i=1}^{N_p} \text{ln}(P(\xi_i|X_i))$$

It is evident that the posterior distribution $P(X|\xi, M_p)$ of $X$ is needed to evaluate $P(\xi|X)$ in Eq. (7). As discussed in the previous section, $P(X|\xi, M_p)$ is numerically represented by $N_p$ MCMCS samples of $X$. Based on these $N_p$ MCMCS samples, $P(\xi|X)$ is estimated as:

$$\text{ln}(P(\xi|X)) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} \text{ln}(P(\xi_i|X_i))$$

in which $P(\xi_i|X_i)$ are the respective values of prior distribution, likelihood function, and posterior distribution evaluated at $i$-th MCMCS sample $x_i$ (i.e., $[\mu, \sigma, \lambda]$) of $X$. $P(\xi|X)$ and $P(X|\xi, M_p)$ are calculated using Eqs. (2) and (4) for each $x_i$, respectively. Evaluating $P(X|\xi, M_p)$ needs the posterior distribution $P(X|\xi, M_p)$ in Eq. (1), which is unknown and is numerically represented by $N_p$ MCMCS samples of $X$ in this study.

This section applies copula theory to re-construct $P(X|\xi, M_p)$ based on MCMCS samples of $X$. Copula theory provides a rational tool to construct multivariate distributions according to multivariate data (e.g., McNeill et al., 2005; Nelsen, 2006; Au and Wang, 2014; Li et al., 2015b; Tang et al., 2015). To our knowledge, this paper is the first contribution that uses copula to describe the multi-dimensional posterior distribution in Bayesian analyses. In the context of copula theory (e.g., Nelsen, 2006), $P(X|\xi, M_p)$ can be written as:

$$P(X|\xi, M_p) = d_{c,\theta}[F(\mu, M_p), F(\sigma, M_p), F(\lambda, M_p); \theta]$$

in which $d_{c,\theta}[-]$ is the copula density function reflecting the dependence among $\mu, \sigma$, and $\lambda$ according to posterior knowledge; $F(\mu, M_p)$, $F(\sigma, M_p)$, and $F(\lambda, M_p)$ are posterior marginal CDFs of $\mu, \sigma$, and $\lambda$, respectively; and $f(\mu, M_p)$, $f(\sigma, M_p)$, and $f(\lambda, M_p)$ are posterior marginal PDFs of $\mu, \sigma$, and $\lambda$, respectively. The posterior marginal CDFs and PDFs of $\mu, \sigma$, and $\lambda$ are estimated by performing conventional statistical analyses on their respective MCMCS samples. Consider, for example, using the Gaussian copula herein. Then, the $d_{c,\theta}[-]$ in Eq. (9) is given by (e.g., Nelsen, 2006):

$$d_{c,\theta}[F(\mu, M_p), F(\sigma, M_p), F(\lambda, M_p); \theta] = |\theta|^{-1} \exp \left( -\frac{1}{2} \phi^\alpha (\theta - \mathbf{E}) \cdot \phi^\beta \right)$$

in which $\phi^\alpha$ is a standard normal random vector with components equal to $\phi^{-1}(F(\mu, M_p))$, $\phi^{-1}(F(\sigma, M_p))$, and $\phi^{-1}(F(\lambda, M_p))$, where $\phi^{-1}(-)$ is the inverse function of standard normal CDF; $\mathbf{E}$ is 3-by-3 identity matrix; and $\theta$ is the copula parameter matrix that describes the correlation coefficient of $\mu, \sigma$, and $\lambda$, and it is written as:

$$\theta = \begin{pmatrix} 1 & \theta_{\alpha\beta} & \theta_{\alpha\lambda} \\ \theta_{\alpha\beta} & 1 & \theta_{\lambda\alpha} \\ \theta_{\alpha\lambda} & \theta_{\lambda\alpha} & 1 \end{pmatrix}$$

in which $\theta_{\alpha\beta}$, $\theta_{\alpha\lambda}$, and $\theta_{\lambda\alpha}$ are the copula parameters describing the correlation coefficient between $\mu$ and $\sigma$ between $\mu$ and $\lambda$, and between $\sigma$ and $\lambda$, respectively. Note that Eq. (11) quantifies the correlation between random field parameters. The absolute values of copula parameters in Eq. (11) reflect whether the random field parameters are correlated or not and how strong their correlations are according to the posterior knowledge. If the copula parameter of two random field parameters is close to zero, the correlation between them is weak in light of the posterior knowledge. On the other hand, if the absolute value of the copula parameter is relatively large (say close to unity for Gaussian Copula), the two random field parameters are strongly correlated based on the posterior knowledge. $\theta_{\alpha\beta}$ is a function of the correlation coefficient between $\mu$ and $\sigma$, and it is calculated as $\sin(\pi \tau_{\alpha\beta}/2)$ in the [0, 1] interval.
this study, where $\tau_{\text{K RCC}}$ is Kendall rank correlation coefficient (KRCC) between $\mu$ and $\sigma$. Similarly, $\theta_{\text{K RCC}}$ and $\theta_{\text{K RCC}}$ are evaluated as $\sin(\pi\text{MCMC}\sigma/2)$ and $\sin(\pi\text{MCMC}\sigma/2)$, respectively, where $\tau_{\text{K RCC}}$ is the KRCC between $\mu$ and $\lambda$, and $\tau_{\text{K RCC}}$ is the KRCC between $\sigma$ and $\lambda$. The KRCCs $\tau_{\text{K RCC}}$, $\tau_{\text{K RCC}}$, and $\tau_{\text{K RCC}}$ are estimated from MCMCS samples of $\mu$, $\sigma$, and $\lambda$:

$$
\tau_{\text{K RCC}} = \frac{1}{0.5N_{p}(N_{p}-1)} \sum_{i,j=1,2,\ldots,N_{p}} \text{sign} \left( \left( \mu_{i} - \mu_{j} \right) \left( \sigma_{i} - \sigma_{j} \right) \right),
$$

$$
\tau_{\text{K RCC}} = \frac{1}{0.5N_{p}(N_{p}-1)} \sum_{i,j=1,2,\ldots,N_{p}} \text{sign} \left( \left( \mu_{i} - \mu_{j} \right) \left( \lambda_{i} - \lambda_{j} \right) \right),
$$

$$
\tau_{\text{K RCC}} = \frac{1}{0.5N_{p}(N_{p}-1)} \sum_{i,j=1,2,\ldots,N_{p}} \text{sign} \left( \left( \sigma_{i} - \sigma_{j} \right) \left( \lambda_{i} - \lambda_{j} \right) \right),
$$

in which $\text{sign}[]$ is the sign function. When $(\mu_{i} - \mu_{j})(\sigma_{i} - \sigma_{j}) > 0$, $\text{sign}[]$ is taken as 1; otherwise, $\text{sign}[]$ is equal to $-1$. Then, substituting Eqs. (9)–(14) into Eq. (8) gives the values of $\ln(P(\xi|M_{k}))$ for $M_{k}$.

For a given $M_{k}$, the MCMCS-based Bayesian approach developed in the previous section is applied to generating random samples of $\mu$, $\sigma$, and $\lambda$ from the posterior distribution. These random samples are subsequently used in Eq. (8) to evaluate $\ln(P(\xi|M_{k}))$. The procedure is repetitively performed $N_{\text{CFT}}$ times for $M_{k}$ ($k = 1, 2, \ldots, N_{\text{CFT}}$) to obtain their respective values of $\ln(P(\xi|M_{k}))$. Then, $M^{*}$ is determined by comparing the values of $\ln(P(\xi|M_{k}))$ of candidate correlation functions. The one with the maximum value of $\ln(P(\xi|M_{k}))$ is taken as $M^{*}$. Its corresponding posterior statistics (e.g., the mean values $\mu'$, $\sigma'$, and $\lambda'$, and standard deviations $\sigma_{\mu}', \sigma_{\sigma}', \sigma_{\lambda}'$, and distributions $\xi^{\mu'}, \xi^{\sigma'}, \xi^{\lambda'}$) are summarized as follows:

1. Obtain a set of cone tip resistance $q_{c}$ versus depth data from CPT test and convert them to the values of $\xi$ (i.e., $\ln(q_{c})$) at different depths $D_{1}$, $D_{2}$, $\ldots$, $D_{n}$ required in Eq. (4), where $q_{c}$ is defined as $(q_{c} - p_{0})/(\sigma_{v0}^{0.5})$. $q_{c}$ is the cone resistance measured at the cone tip, $\sigma_{v0}$ is the vertical effective stress, and $p_{0}$ is the standard atmospheric pressure (i.e., 100 kPa);
2. Obtain a set of prior knowledge on $\mu$, $\sigma$, and $\lambda$, such as the reasonable ranges of $\mu$, $\sigma$, and $\lambda$ with their respective minimum values $\mu_{\text{min}}$, $\sigma_{\text{min}}$, and $\lambda_{\text{min}}$, and respective maximum values $\mu_{\text{max}}$, $\sigma_{\text{max}}$, and $\lambda_{\text{max}}$ which are needed in Eq. (2);
3. Determine $N_{\text{CFT}}$ candidate correlation functions $M_{k}$ ($k = 1, 2, \ldots, N_{\text{CFT}}$), such as SECF, BNCF, SMCF, and SQCF shown in Fig. 1;
4. Use M-H algorithm to generate $N_{s}$ samples of $\mu$, $\sigma$, and $\lambda$ based on Eqs. (1), (2), and (4) for a given $M_{k}$;

**Fig. 3** shows a flow chart for the implementation of the proposed Bayesian approaches schematically. In general, the implementation procedure involves 8 steps. Details of each step and its associated equations are summarized as follows:

1. Obtain a set of cone tip resistance $q_{c}$ versus depth data from CPT test and convert them to the values of $\xi$ (i.e., $\ln(q_{c})$) at different depths $D_{1}$, $D_{2}$, $\ldots$, $D_{n}$ required in Eq. (4), where $q_{c}$ is defined as $(q_{c} - p_{0})/(\sigma_{v0}^{0.5})$. $q_{c}$ is the cone resistance measured at the cone tip, $\sigma_{v0}$ is the vertical effective stress, and $p_{0}$ is the standard atmospheric pressure (i.e., 100 kPa);
2. Obtain a set of prior knowledge on $\mu$, $\sigma$, and $\lambda$, such as the reasonable ranges of $\mu$, $\sigma$, and $\lambda$ with their respective minimum values $\mu_{\text{min}}$, $\sigma_{\text{min}}$, and $\lambda_{\text{min}}$, and respective maximum values $\mu_{\text{max}}$, $\sigma_{\text{max}}$, and $\lambda_{\text{max}}$ which are needed in Eq. (2);
3. Determine $N_{\text{CFT}}$ candidate correlation functions $M_{k}$ ($k = 1, 2, \ldots, N_{\text{CFT}}$), such as SECF, BNCF, SMCF, and SQCF shown in Fig. 1;
4. Use M-H algorithm to generate $N_{s}$ samples of $\mu$, $\sigma$, and $\lambda$ based on Eqs. (1), (2), and (4) for a given $M_{k}$;
(5) Estimate the respective posterior statistics (e.g., \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\)) and distributions (e.g., PDFs and CDFs) of \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\) based on their MCMCS samples through conventional statistical analyses;

(6) Calculate \(\ln (P(\xi, s_r, s_\alpha | M_k))\) of \(M_k\) using Eqs. (8)–(14), and the evidence \(P(\xi, s_r, s_\alpha | M_k)\) is then obtained by taking the exponential of \(\ln (P(\xi, s_r, s_\alpha | M_k))\);

(7) Repeat steps (4)–(6) \(N_C\) times to generate \(N_M\) MCMCS samples of \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\) for each candidate correlation function;

(8) Compare the values of \(\ln (P(\xi, s_r, s_\alpha | M_k))\) (or \(P(\xi, s_r, s_\alpha | M_k)\)) of the \(N_C\) candidate correlation functions and select the one with the maximum value of \(\ln (P(\xi, s_r, s_\alpha | M_k))\) as \(M^*\). Its corresponding posterior statistics and distributions reflect the posterior knowledge on \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\) and probabilistically characterize the ISV of \(\varphi^*\).

These eight steps can be readily programmed as a user function in commonly available commercial software packages, such as MATLAB (e.g., Mathworks Inc., 2015). A MATLAB user-function is developed in this study to implement the proposed Bayesian approaches. The corresponding M-file is provided as supplementary data in this paper. Interested readers can also contact the corresponding author for such a user function.

For illustration, the proposed Bayesian approaches are applied to probabilistically characterize the ISV of \(\varphi^*\) at a sand site of NGES at Texas A&M University (e.g., Briaud, 1997). The site is comprised of a top layer of sand to about 12 m and underlain by hard clay thereafter. The groundwater table is about 6 m deep, and the total unit weight of sand is about 18.4 kN/m². Fig. 4 shows a set of cone tip resistance \(q_c\) versus depth obtained from CPT at this site, in which the average depth interval between \(q_c\) data is around 0.07 m. The set of CPT data is used as input in the proposed Bayesian approaches to determine the random field parameters and select an appropriate correlation function among a pool of candidates for probabilistic characterization of ISV of \(\varphi^*\) at the sand site.

Consider, for example, four candidate correlation functions shown in Fig. 1, i.e., SECF, BNCF, SMCF, and SQECF, which are denoted as \(M_1, M_2, M_3, \text{and } M_4\) in this study, respectively. The prior knowledge on \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\) is taken as a joint uniform distribution given by Eq. (2), in which the typical ranges of \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\) are taken as \([20°, 40°], [1°, 6°], [0 m, 6 m], [20°, 40°], [1°, 6°]\), and \([0 m, 6 m]\), respectively. These typical ranges are consistent with those reported in literature (e.g., Phoon and Kulhawy, 1999a). Based on the set of CPT data shown in Fig. 4 and the prior knowledge, a MCMCS run is performed to generate 100,000 samples of \(\mu, \sigma, \lambda, \xi, s_r, s_\alpha, \tau_{\text{MN}}, \tau_{\text{MP}}, \text{and } \tau_{\text{GT}}\) from the posterior distribution for each candidate correlation function, the number of which is sufficient for MCMCS to reach the stationary condition of Markov Chain and to generate reasonably accurate
Fig. 9. Posterior marginal CDFs and PDFs of $\mu$, $\sigma$ and $\lambda$ for $M_1$ estimated from MCMCS samples: (a), (b), (c) Posterior marginal CDFs of $\mu$, $\sigma$ and $\lambda$; (d), (e), (f) Posterior marginal PDFs of $\mu$, $\sigma$ and $\lambda$.

Fig. 10. Posterior marginal CDFs and PDFs of $\mu$, $\sigma$ and $\lambda$ for $M_2$ estimated from MCMCS samples: (a), (b), (c) Posterior marginal CDFs of $\mu$, $\sigma$ and $\lambda$; (d), (e), (f) Posterior marginal PDFs of $\mu$, $\sigma$ and $\lambda$. 
Fig. 11. Posterior marginal CDFs and PDFs of $\mu$, $\sigma$ and $\lambda$ for $M_3$ estimated from MCMCS samples: (a), (b), (c) Posterior marginal CDFs of $\mu$, $\sigma$ and $\lambda$; (d), (e), (f) Posterior marginal PDFs of $\mu$, $\sigma$ and $\lambda$.

Fig. 12. Posterior marginal CDFs and PDFs of $\mu$, $\sigma$ and $\lambda$ for $M_4$ estimated from MCMCS samples: (a), (b), (c) Posterior marginal CDFs of $\mu$, $\sigma$ and $\lambda$; (d), (e), (f) Posterior marginal PDFs of $\mu$, $\sigma$ and $\lambda$. 
probabilistic estimates of $\mu$, $\sigma$, and $\lambda$ in this study. Note that the efficiency of MCMCS depends on the choice of the initial sample of Markov Chain and the proposal PDF that is needed in the simulation (e.g., Zhang et al., 2010). In this example, the initial samples of the Markov Chain are taken as the prior mean values (i.e., $(\mu_{\text{max}} + \mu_{\text{min}})/2$, $(\sigma_{\text{max}} + \sigma_{\text{min}})/2$, and $(\lambda_{\text{max}} + \lambda_{\text{min}})/2$ for uniform prior distributions) of $\mu$, $\sigma$, and $\lambda$, respectively. The proposal PDF is defined as a joint normal PDF of $\mu$, $\sigma$, and $\lambda$ that is centered at the previous state of Markov Chain and has a covariance matrix $\Sigma_{\text{MCMC}}$. Herein, $\Sigma_{\text{MCMC}}$ is simply taken as a 3-by-3 diagonal matrix with main diagonal elements equal to the product of the previous state and prior coefficients of variation (i.e., $(\mu_{\text{max}} - \mu_{\text{min}})/\sqrt{3}$, $(\sigma_{\text{max}} - \sigma_{\text{min}})/\sqrt{3}$, and $(\lambda_{\text{max}} - \lambda_{\text{min}})/\sqrt{3}$ for uniform prior distributions) of $\mu$, $\sigma$, and $\lambda$. Although a large number (i.e., 100,000) of MCMCS samples are generated in this example, a MCMCS run is terminated when the acceptance ratio is less than 0.2. In other words, a large number of MCMCS samples can be generated with relative ease in this study. In such a case, effects of the initial sample and the proposal PDF on the efficiency of MCMCS become less critical.

5.1. MCMCS samples of random field parameters

Figs. 5–8 show scatter plots of 100,000 MCMCS samples of $\mu$, $\sigma$, and $\lambda$ for $M_1$–$M_4$, respectively. Based on these MCMCS samples, KRCs (i.e., $k_{\mu\mu}$, $k_{\sigma\sigma}$, and $k_{\mu\sigma}$) between them and the corresponding Gaussian copula parameters (i.e., $\theta_{\mu\mu}$, $\theta_{\sigma\sigma}$, and $\theta_{\mu\sigma}$) are estimated, which are also included in Figs. 5–8. Figs. 5(a) and (b) show that $\mu$ is weakly correlated with $\sigma$ and $\lambda$ in a negative manner for $M_1$, while a relatively strong positive correlation between $\sigma$ and $\lambda$ is observed in Fig. 5(c) for $M_1$. Similar observations can also be obtained from Figs. 6–8 for $M_2$–$M_4$, respectively. In addition, conventional statistical analyses are also performed on the MCMCS samples shown in Figs. 5–8 to obtain posterior statistics (e.g., mean values and standard deviations) and distributions (e.g., marginal PDFs and CDFs) of $\mu$, $\sigma$, and $\lambda$.

Figs. 9–12 plot the posterior marginal PDFs and CDFs of $\mu$, $\sigma$, and $\lambda$ estimated from 100,000 MCMCS samples generated using $M_1$–$M_4$ by solid lines, respectively. For validation, direct numerical integration is also performed to calculate the posterior marginal PDFs and CDFs of $\mu$, $\sigma$, and $\lambda$ using Eq. (1). In the direct numerical integration, the three-dimensional (3D) space of $\mu$, $\sigma$, and $\lambda$ is discretized into a number of 3D-cubes, for each of which likelihood function and prior distribution are evaluated at the centroid and the volume of the cube is calculated as well. Then, products of the likelihood function, prior distribution, and volume for each 3D-cube are calculated, which sum up to be the normalizing constant $K$ in Eq. (1). By this means, $K$ is calculated numerically and repeatedly for $M_1$–$M_4$ over a 3D space of $\mu$, $\sigma$, and $\lambda$ to obtain the posterior distribution (i.e., Eq. (1)), in which the Gaussian copula is not used. Then, the posterior marginal distribution of one parameter (e.g., $\lambda$) is evaluated by integrating Eq. (1) over the space of the other two parameters (e.g., $\mu$ and $\sigma$). Figs. 9–12 also include the posterior marginal PDFs and CDFs of $\mu$, $\sigma$, and $\lambda$ obtained from direct numerical integration for $M_1$–$M_4$ by open triangles, respectively. For a given correlation function, results obtained from the two approaches (i.e., MCMCS (see solid lines) and direct numerical integration (see open triangles)) are generally in good agreement. This indicates that 100,000 MCMCS samples of $\mu$, $\lambda$, and $\sigma$ portray the posterior distribution reasonably well for $M_1$–$M_4$ in this example.

Generally speaking, the posterior marginal PDFs of $\mu$ and $\sigma$ for $M_1$–$M_4$ peak at a single value or around a narrow proximity, as shown in Figs. 9(d)–(e), 10(d)–(e), 11(d)–(e), and 12(d)–(e). This indicates that $\mu$ and $\sigma$ are globally identifiable in this example no matter which correlation function is used to interpret the CPT data. On the other hand, different behaviors of the posterior marginal PDFs of $\lambda$ are observed for $M_1$–$M_4$, as shown in Figs. 9(f), 10(f), 11(f), and 12(f). For $M_4$, the posterior marginal PDF of $\lambda$ varies slightly as $\lambda$ is less than 2 m (see Fig. 9(f)), indicating that $\lambda$ is likely to be unidentifiable in this example as $M_4$ is used. For $M_2$, a multi-modal nature of posterior marginal PDF of $\lambda$ is observed (see Fig. 10(f)), indicating that $\lambda$ might be locally identifiable in this example when $M_2$ is used to interpret the CPT data. For $M_3$ and $M_4$, the posterior marginal PDF of $\lambda$ peaks at a single value (see Figs. 11(f) and 12(f)), which means that $\lambda$ is globally identifiable in this example when $M_3$ or $M_4$ is adopted. It is evident that, for a given set of CPT data and prior knowledge, the adopted correlation function affects the feature of posterior distribution of random field parameters. As a result, the random field parameters (e.g., $\lambda$ in this example) of soil properties can be globally identifiable, locally identifiable, or unidentifiable, particularly in the case with a limited number of project-specific test data. This situation has led to difficulty in some approximate methods in Bayesian analysis, such as LAAM, since it is only valid in globally identifiable cases. Further discussion is given in a later subsection. In contrast, MCMCS provides a rational tool to solve the posterior distribution of random field parameters for probabilistic characterization of ISV no matter which correlation function is applied.

Table 1 Results of random field parameters and correlation function obtained from the proposed Bayesian approaches.

| Correlation function | Logarithm of the evidence $\ln(P(\mathbf{\xi}|\mathbf{M}_i))$ | $P(M_i|\mathbf{\xi})$ | Posterior mean value $\mu$ (°) | $\sigma$ (°) | $\lambda$ (m) | Posterior standard deviation $\sigma_\mu$ (°) | $\sigma_\sigma$ (°) | $\sigma_\lambda$ (m) |
|----------------------|---------------------------------|-------------------|------------------|--------|--------|-----------------|----------------|--------|
| $M_1$: SECF         | -116.83                         | 0.193             | 38.3             | 3.7    | 3.7    | 1.3             | 0.8            | 1.4    |
| $M_2$: BNCF         | -117.29                         | 0.124             | 38.2             | 4.3    | 3.4    | 1.4             | 1.0            | 1.2    |
| $M_3$: SMCF$^a$     | -116.03                         | 0.440             | 38.5             | 3.8    | 1.7    | 1.1             | 0.9            | 0.8    |
| $M_4$: SQCF         | -116.60                         | 0.243             | 38.7             | 3.5    | 1.0    | 0.9             | 0.8            | 0.4    |

$^a$ Most probable correlation function.

Table 2 Results of random field parameters and correlation function obtained from direct numerical integration.

| Correlation function | Logarithm of the evidence $\ln(P(\mathbf{\xi}|\mathbf{M}_i))$ | $P(M_i|\mathbf{\xi})$ | Posterior mean value $\mu$ (°) | $\sigma$ (°) | $\lambda$ (m) | Posterior standard deviation $\sigma_\mu$ (°) | $\sigma_\sigma$ (°) | $\sigma_\lambda$ (m) |
|----------------------|---------------------------------|-------------------|------------------|--------|--------|-----------------|----------------|--------|
| $M_1$: SECF         | -116.93                         | 0.191             | 38.3             | 3.7    | 3.7    | 1.3             | 0.8            | 1.3    |
| $M_2$: BNCF         | -117.36                         | 0.121             | 38.2             | 4.3    | 3.4    | 1.4             | 1.0            | 1.2    |
| $M_3$: SMCF$^a$     | -116.07                         | 0.440             | 38.5             | 3.7    | 1.7    | 1.1             | 0.9            | 0.8    |
| $M_4$: SQCF         | -116.65                         | 0.248             | 38.7             | 3.5    | 0.5    | 0.9             | 0.8            | 0.4    |

$^a$ Most probable correlation function.
Based on the KRCCs (or Gaussian copula parameters) shown in Figs. 5–8 and posterior marginal distributions shown in Figs. 9–12, \( \ln(P(q|M_i)) \) for each candidate correlation function is calculated using Eqs. (8)–(14) for determining \( M^* \) among \( M_1-M_4 \), as discussed in the next subsection.

### 5.2. Most probable correlation function and its corresponding posterior knowledge

Table 1 summarizes the values of \( \ln(P(q|M_i)) \) and occurrence probabilities (i.e., \( P(q|M_i) \)) for \( M_1-M_4 \) in Columns 2 and 3. It is shown that \( M_2 \) (i.e., SMCF) has the largest value (i.e., \(-116.03\)) of \( \ln(P(q|M_i)) \) among \( M_1-M_4 \) in this example. Therefore, it is taken as \( M^* \) for modeling ISV of \( \phi^e \) at the sand site of NGES, and its corresponding occurrence probability is about 0.44 for the given set of CPT data shown in Fig. 4 and the uniform prior knowledge adopted in this example. In addition, Table 1 also includes the posterior statistics (e.g., \( \mu', \sigma', \lambda', s_{\mu'}, s_{\sigma'}, s_{\lambda'} \)) of \( M_1-M_4 \), which are estimated from MCMCS samples shown in Figs. 5–8. It is evident that the posterior statistics of random field parameters depend on the correlation function adopted in the analysis. For example, the posterior mean values (i.e., 3.7 m and 3.4 m) of \( \lambda \) for \( M_1 \) and \( M_2 \) are much greater than those (i.e., 1.7 m and 1.0 m) for \( M_3 \) and \( M_4 \) (see Column 6 in Table 1). Selection of a proper correlation function is necessary prerequisite for determining the posterior statistics of random field parameters for probabilistic characterization of ISV. In this example, the posterior statistics corresponding to \( M_2 \) (i.e., \( M^* \)) reflect the posterior knowledge on random field parameters (i.e., \( \mu, \sigma \), and \( \lambda \)) of \( \phi^e \) at the sand site. Effects of the correlation function on random field parameters are rationally taken into account in the proposed Bayesian approaches.

For validation, the logarithm of evidence, occurrence probability, and posterior statistics for each candidate correlation function are also calculated by performing direct numerical integration on Eqs. (1) and (6), as summarized in Table 2. The results obtained from direct numerical integration (see Table 2) agree well with those obtained from the proposed approaches using MCMCS and Gaussian Copula (see Table 1). This validates the proposed approaches.

### 5.3. Comparison with LAAM

For comparison, LAAM is also applied to obtaining the posterior knowledge of random field parameters for each candidate correlation function.
function and to evaluating the evidence. Using LAAM, the posterior distribution (e.g., Eq. (1)) is approximated by a Gaussian distribution. Then, the posterior mean values of random field parameters are equal to their respective most probable values (MPVs) (i.e., $\mu^*$, $\sigma^*$, and $\lambda^*$) that can be obtained by maximizing the posterior distribution $P(X|\hat{\xi}, M)$ (or equivalently minimizing the $-\ln(P(X|\hat{\xi}, M))$ in implementation). In addition, the posterior standard deviations are calculated through evaluating the inverse of Hessian matrix of $\ln(P(X|\hat{\xi}, M))$ at the posterior MPVs. More details of using LAAM to obtain posterior knowledge and to evaluate the evidence for model comparison are referred to Yuen (2010a, 2010b) and Cao and Wang (2013).

Table 3 shows the results estimated from LAAM, including the logarithm of evidence, occurrence probability, and posterior statistics for each candidate correlation function. Based on the values of evidence estimated from LAAM, $M_2$ (i.e., SMCF) is also selected as $M^*$ among $M_1$–$M_4$ in this example, and its corresponding occurrence probability is 0.432, which are consistent with those obtained from the proposed approaches. However, it should be noted that the occurrence probability (i.e., 0.027) of $M_2$ estimated from LAAM is about five times smaller than those (i.e., 0.124 and 0.121, respectively) obtained from the proposed approaches and direct numerical integration. Such a difference might be attributed to the multi-modal feature (see Figs. 10(f)) of the posterior distribution for $M_2$. This feature violates the Gaussian approximation adopted in LAAM and leads to incorrect outcomes of $M_2$ by LAAM, although LAAM gives the correct most probable correlation function in this example.

In addition, it is also shown that the posterior knowledge (including mean values or MPVs) and standard deviations, as shown in Columns 4–9 in Table 3) of $\mu$, $\sigma$ and $\lambda$ obtained from LAAM are considerably different from those (see columns 4–9 in Table 1) estimated from MCMCS samples. This, again, indicates that the posterior distribution $P(X|\hat{\xi}, M)$ in Eq. (1) may not be well approximated by a Gaussian distribution in this example, particularly when $M_1$ or $M_2$ is used in Eq. (1), with which $\lambda$ is probably unidentifiable or locally identifiable (see Figs. 9(f) and 10(f)). In such a case, MCMCS provides a more general tool to solve the posterior knowledge of random field parameters for probabilistic characterization of ISV of $\varphi$.

6. Validation using simulated CPT data

Note that in practice the actual values of soil properties are unknown, and they are estimated through prior knowledge and project-specific test results. For further validation, the proposed Bayesian approaches are used to identify random field parameters and to select the most probable correlation function together with simulated CPT data in this section, in which true values of random field parameters and the actual correlation function are known. The CPT data are simulated from Eq. (4) with $\mu$, $\sigma$, and $\lambda$ equal to 30°, 3°, and 0.5 m, respectively. As indicated by Eq. (4), a spatial correlation function is needed to define the spatial correlation of $\varphi$ in the simulation. Consider the four correlation functions shown in Fig. 1 (i.e., SECF, BNCF, SMCF, and SQCF, denoted as $M_1$, $M_2$, $M_3$, and $M_4$ respectively). For each correlation function, 10 sets of $\xi$ (i.e., $\ln(q)$) are simulated for a penetration depth up to 50 m (i.e., 100$\lambda$) at a depth interval of 0.05 m (i.e., $\lambda/10$), resulting in a total of 40 sets of simulated $\ln(q)$ data. For example, Figs. 13(a)–(d) show four sets of $\ln(q)$ data simulated using $M_1$, $M_2$, $M_3$, and $M_4$ in Eq. (4), respectively. Using each set of CPT data and the prior knowledge adopted in the previous section, the proposed Bayesian approaches provide the values of $\ln(P(\hat{\xi}|M_j))$ of candidate correlation functions (e.g., $M_1$–$M_4$) for determining $M^*$, and the posterior estimates (e.g., mean values) of $\mu$, $\sigma$, and $\lambda$ for each candidate correlation function. Note that the accuracy of results obtained from Bayesian analyses relies on the amount of test data used in the analysis (e.g., Wang and Cao, 2013; Cao and Wang, 2014a). The amount of CPT data depends on the sampling depth and

![Fig. 14. Comparison of the evidence estimated from MCMCS samples using Gaussian copula and direct numerical integration.](image-url)
interval. In this section, a relatively long sampling depth (i.e., 100λ) and small sampling interval (i.e., λ/10) are adopted to ensure that the simulated CPT data contains sufficient information on the true random field parameters and the actual correlation function used in the simulation. The relationship between data characteristics (e.g., sampling depth and interval) of CPT profile and accuracy of Bayesian characterization of ISV is worthwhile to be systematically explored in future study.

For comparison, the \( M' \), \( \mu \), \( \sigma \), and \( \lambda \) are directly estimated from CPT data using the regression shown in Fig. 2. For each set of CPT data, a profile of \( \phi' \) versus depth is directly calculated from the regression shown in Fig. 2. Then, conventional statistical analyses are performed on the profile of \( \phi' \) to estimate its \( \mu \), \( \sigma \), and sample autocorrelation function (SACF). Then, the \( \lambda \) and \( M' \) are determined by fitting a candidate correlation function (e.g., \( \phi_1 - \phi_4 \)) to the SACF and comparing the goodness-of-fitting of different candidate correlation functions (e.g., Uzielli et al., 2005; Stuedlein et al., 2012). The above procedures result in a set of estimates of \( \mu \), \( \sigma \), \( \lambda \), and \( M' \) for each set of simulated CPT data, which are compared with results obtained from the proposed Bayesian approaches.

6.1. Validation of Bayesian model selection results

Fig. 14 plots the values of \( \ln(P(\xi|\phi(M_1))) \) estimated from the proposed approach using MCMCS and Gaussian copula versus those obtained from direct numerical integration for the 40 sets of simulated CPT data. For each set of CPT data, the values of \( \ln(P(\xi|\phi(M_1))) \) for \( \phi_1 - \phi_4 \) are calculated using the proposed approach and direct numerical integration, respectively. The results obtained from the two approaches agree well with each other, indicating that the proposed approach calculates the evidence properly based on MCMCS samples and Gaussian copula. Table 4 summarizes the results of \( M' \) estimated from 40 sets of simulated CPT data using the proposed approach. For the 10 sets of CPT data simulated using \( \phi_1 \) (i.e., SECF), the most probable correlation functions obtained from the proposed approach are all identical to the true correlation function \( \phi_1 \). For the respective 10 sets of CPT data simulated using \( \phi_2 - \phi_4 \), 8 sets of them give the correct correlation function, including in the true correlation function \( \phi_i \). In total, 34 out of 40 sets of simulated CPT data in the proposed approach identifies the correct correlation function, and the rate of correct identification is 85% using the proposed approach. The remaining 6 sets of simulated CPT data give incorrect correlation functions. This might be attributed to the random fluctuation in the simulated CPT data. Such random fluctuation renders these 6 sets of simulated CPT data less similar to their corresponding true correlation functions adopted in simulation than other correlation functions.

For comparison, Table 4 also summarizes the results of correlation functions selected by comparing the goodness-of-fitting of \( \phi_1 - \phi_4 \) with the SACF of the profile of \( \phi' \), which is estimated from CPT data using the regression model, in the parentheses. The results show that all the 40 sets of simulated data suggest that \( M' \) is the SECF (i.e., \( \phi_1 \)) without regard to the true correlation function used in simulation, most of which are obviously incorrect except the results from CPT data simulated using \( \phi_1 \). Note that the CPT data \( \xi \) is simulated using Eq. (4). As indicated by the covariance matrix \( \Sigma (i.e., \sigma^2 \xi^2 + \sigma^2 \phi) \) of \( \xi \) in Eq. (4), the fluctuation of simulated CPT data not only relies on the ISV (i.e., \( \sigma^2 \xi \)) of \( \phi' \), which is specified by the true correlation function used in simulation, but also depends on the transformation uncertainty (i.e., \( \sigma^2 \phi \)) associated with the regression model. Both the ISV of \( \phi' \) and the transformation uncertainty propagate into the profile of \( \phi' \) (e.g., those shown in Figs. 13(e)–(h)) directly estimated from the CPT data using the regression model. Therefore, the SACF estimated from \( \phi' \) profile does not reflect the actual spatial correlation of \( \phi' \) specified by the true correlation function in simulation. It is then, not surprising to see that there is a high chance (e.g., 30 out of 40 sets of simulated data in this
6.2. Validation of posterior estimates of random field parameters

Fig. 15(a), (b), and (c) show posterior mean values of \( \mu, \sigma \), and \( \lambda \) obtained from the proposed Bayesian approaches using the 40 sets of simulated CPT data by circles, respectively. For comparison, Fig. 15 also includes the true values of \( \mu, \sigma \), and \( \lambda \) by solid lines. It is shown that all the circles plot close to the solid lines, indicating that the proposed Bayesian approaches properly identify random field parameters for probabilistic characterization of ISV. In addition, Fig. 15 also shows the values of \( \mu \) values (i.e., triangles in Fig. 15(a)) from direct estimation compare favorably with the true value, i.e., 30°. However, the \( \sigma \) values (i.e., triangles in Fig. 15(b)) from direct estimation are greater than its true value (i.e., 3°), and the \( \lambda \) values (i.e., triangles in Fig. 15(c)) from direct estimation are less than its true value (i.e., 0.5 m). As discussed in the previous subsection, the variability in the \( \sigma \) values directly estimated CPT data using the regression model contains not only the ISV of \( \phi' \) but also the transformation uncertainty, and it is, therefore, greater that the actual ISV of \( \phi' \) prescribed in simulation, which subsequently leads to overestimation in \( \sigma \) and underestimation in \( \lambda \). Compared with direct estimates from CPT data using the regression model, the proposed Bayesian approaches deal, rationally and transparently, with various uncertainties (e.g., ISV and transformation uncertainty) and provide proper posterior knowledge on the ISV of \( \phi' \) using both CPT data and prior knowledge.

7. Summary and conclusions

This paper developed Bayesian approaches for probabilistic characterization of inherent spatial variability (ISV) of sand effective friction angle \( \phi' \) in a statistically homogeneous sand layer using indirect test data (e.g., cone penetration test (CPT) data) and prior knowledge. The proposed approaches identify random field parameters (i.e., mean \( \mu \), standard deviation \( \sigma \), and scale of fluctuation \( \lambda \)) of the design soil property concerned and simultaneously select the most probable correlation function \( M^* \) among a pool of candidate correlation functions using project-specific CPT data and prior knowledge. For a given correlation function, the information from CPT data and prior knowledge is systematically integrated as the posterior knowledge on \( \mu, \sigma, \) and \( \lambda \), which is quantified by the posterior distribution under a Bayesian framework. Then, a simple Markov Chain Monte Carlo Simulation (MCMCSA) algorithm, i.e., Metropolis-Hastings (M-H) algorithm, is used to generate a large number of \( \mu, \sigma, \) and \( \lambda \) samples from the posterior distribution for its numerical representation. Based on these MCMCSA samples, a Gaussian Copula-based method is proposed to calculate the evidence of candidate correlation functions for selecting \( M^* \). This makes M-H algorithm feasible in Bayesian model selection problems and removes one key limitation of the algorithm.

Equations were derived for the proposed Bayesian approaches, and they were illustrated using a set of real-life CPT data obtained from a sand site of NGES at Texas A&M University. The results (including posterior distributions of \( \mu, \sigma \), and \( \lambda \) of \( \phi' \), and the evidence of candidate correlation functions) obtained from the proposed approaches were validated against those obtained from direct numerical integration. The results obtained from the two approaches agree well with each other. This validates the proposed approaches. It is also found that the posterior estimates of random field parameters depend on the correlation function adopted to interpret CPT data in Bayesian analysis. Effects of the correlation function on random field parameters are rationally taken into account in the proposed approaches. With a limited number of project-specific CPT data, the random field parameters can be globally identifiable, locally identifiable, and unidentifiable. This situation leads to difficulty in some approximation methods (such as Laplace asymptotic approximation method that is only valid for globally identifiable cases) and highlights the suitability of MCMCSA in Bayesian characterization of geotechnical properties.

The proposed approaches were also validated using simulated CPT data, where the true values of \( \mu, \sigma, \) and \( \lambda \) of \( \phi' \) and the correct correlation function are known. It was shown that the probabilistic estimates (e.g., posterior mean values) of \( \mu, \sigma, \) and \( \lambda \) obtained from the proposed approaches agree well with their respective true values, and the most probable correlation function is properly selected using CPT data and prior knowledge in a rational manner. This further validates the proposed approaches. The results were also compared with those obtained from conventional statistical analyses on the \( \phi' \) profile estimated from CPT data using the regression model shown in Fig. 2, which was developed by Kulhawy and Mayne (1990) and empirically links \( \phi' \) to normalized cone tip resistance. It was found that the proposed Bayesian approaches provide more consistent probabilistic characterization (including posterior estimates of \( \mu, \sigma, \) and \( \lambda \), and the most probable correlation function) of the ISV of \( \phi' \) with proper consideration of the transformation uncertainty.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (Project No. 51225903), the National Basic Research Program of China (973 Program) (Project No. 2011CB013502), the National Natural Science Foundation of China (Project Nos. 51329901, 51409196, 51579190, 51528901), and the Natural Science Foundation of Hubei Province of China (Project No. 2014CF001).

References


