Proc. of the 8th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Jinsong Huang, D.V. Griffiths, Shui-Hua Jiang, Anna Giacomini and Richard Kelly

doi: 10.3850/978-981-18-5182-7_01-005-cd

ISGSR 2022

Probability Density Function and Credible Region Estimation for Multivariate Uncertain Irregular Geotechnical Data

Zi-Tong Zhao[1], He-Qing Mu[2,3*], Ka-Veng Yuen[1,4*]

[1]State Key Laboratory of Internet of Things for Smart City and Department of Civil and Environmental Engineering, University of Macau, Macau 999078, P.R. China.
[2]School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510640, P.R. China.
[3]State Key Laboratory of Subtropical Building Science, South China University of Technology, Guangzhou 510640, P.R. China.
[4]Guangdong–Hong Kong-Macau Joint Laboratory for Smart Cities, P.R. China.
Email: yc17401@umac.mo (Z.-T. Zhao), cthqmu@scut.edu.cn (H.-Q. Mu*), kvyuen@um.edu.mo (K.-V. Yuen*)

Abstract: Probability Density Function (PDF) modelling and Credible Region (CR) construction are two key issues or describing Multivariate Uncertain Irregular (MUI) characteristics of geotechnical data. There are two fundamental difficulties in this task. The first is on the joint PDF modelling of complex MUI characteristics, including modelling asymmetry and/or multimodality. The second is on the CR construction of the asymmetric and/or multimodal PDF. These issues have seldom been addressed as these problems were usually considered in the case of asymmetry and unimodality only. Aiming to resolve these two difficulties, this paper proposes BAyeSIan Copula-based Highest posterior density Regions (BASIC-HR). This framework contains Stage-PDF and Stage-CR. Stage-PDF fuses copula theory and Bayesian inference to obtain the posterior joint CDF and PDF. Stage-CR fuses the posterior PDF and the concept of highest density regions to construct the Highest posterior density Regions (HR). The HR is defined as that, given a cumulative probability, any point inside the HR should have posterior probability density at least as large as any point outside the HR. Based on this definition of the HR, CR construction can be performed for MUI geotechnical data. An example is presented to illustrate the capability of the proposed framework.

Keywords: Bayesian Inference; Copula; Credible Region; Multivariate Geotechnical Data.

1 Introduction

Geotechnical data are typically Multivariate Uncertain Irregular (MUI). It is multivariate because multiple tests (e.g., triaxial test, shear strength test) are usually conducted at the same location during site investigation (Ching and Phoon 2020). It is uncertain (Wang et al. 2013, 2016) due to measurement errors, statistical uncertainty, and transformation uncertainty arising from indirect measurement, etc. It is irregular because asymmetry and/or multimodality are frequently observed in data histogram.

For describing MUI characteristics of geotechnical data, Probability Density Function (PDF) modelling and Credible Region (CR) construction are two key issues. There are two fundamental difficulties. The first is on the joint PDF modelling of complex MUI characteristics, including modelling asymmetry and/or multimodality, which makes the collapse of the traditional multivariate distributions. The second is on the CR construction of the asymmetric and/or multimodal PDF. Credible Region (CR) construction is critical in geotechnical design. CR construction of the case of a symmetry and unimodality PDF is well studies, but that of the case of an asymmetric and and/or multimodal PDF is seldomly addressed. Nevertheless, the later case is more common in the design based on MUI geotechnical data.

Aiming to resolve these two difficulties, this paper proposes BAyeSIan Copula-based Highest posterior density Regions (BASIC-HR). This framework contains Stage-PDF and Stage-CR. Stage-PDF fuses copula theory and Bayesian inference to obtain the posterior joint CDF and PDF. Stage-CR fuses the posterior PDF and the concept of Highest posterior density Regions (HR). The HR is defined as that, given a cumulative probability, any point inside the HR should have posterior probability density at least as large as any point outside the HR (Box and Tiao 2011).

The structure of this paper is outlined as follows. Section 2 proposes the two-stage BASIC-HR. Section 3 are illustrative examples.

2 Proposed Framework

2.1 Stage-PDF

Copula can be regarded as a function that connects the multivariate joint PDF of Random Variables (RVs) with its univariate marginal PDFs (Nelsen 2007; Sklar 1996, 1959). Use $\boldsymbol{X} = [X_1, X_2, \cdots, X_D]$ to denote the set of D-

dimensional RVs, where X_d denote the d-th univariate RV. Copula is defined as a D-dimensional joint CDF whose univariate marginal CDFs on the unit hypercube $[0,1]^D$ is the standard uniform distribution on the interval $[0,1]$ (Nelsen 2007). Let $F_d(x_d)$ and $p_d(x_d)$ denote the univariate marginal CDF and PDF of the d-th univariate RV, respectively. Let $F_{1:D}(x_1, x_2, \cdots, x_D)$ denote the D-dimensional joint CDF with univariate marginal CDFs $F_1(x_1), F_2(x_2), \cdots, F_D(x_D)$. Sklar's theorem states that there exists a Copula $C_{1:D}(u_1, u_2, \cdots, u_D)$ connecting the D-dimensional copula-based joint CDF with its univariate marginal CDFs (Nelsen 2007; Sklar 1996, 1959):

$$F_{1:D}(x_1, x_2, \cdots, x_D) = C_{1:D}\big(F_1(x_1), F_2(x_2), \cdots, F_D(x_D)\big) = C_{1:D}(u_1, u_2, \cdots, u_D) \tag{1}$$

A copula of $(X_1, X_2, \cdots, X_D)$ is defined as the joint CDF of $(U_1, U_2, \cdots, U_D)$:

$$C_{1:D}(u_1, u_2, \cdots, u_D) = P(U_1 \le u_1, U_2 \le u_2, \cdots, U_D \le u_D) \tag{2}$$

The joint copula-based PDF $p_{1:D}(x_1, x_2, \cdots, x_D)$ of RVs $\boldsymbol{X}$ can be obtained from its joint CDF $F_{1:D}(x_1, x_2, \cdots, x_D)$ of Eq. (1):

$$p_{1:D}(x_1, x_2, \cdots, x_D) = \frac{\partial^D p_{1:D}(x_1, x_2, \cdots, x_D)}{\partial x_1 \partial x_2 \cdots \partial x_D} = \frac{\partial^D C_{1:D}(u_1, u_2, \cdots, u_D)}{\partial u_1 \partial u_2 \cdots \partial u_D} \prod_{i=1}^{D} \frac{\partial F_i(x_i)}{\partial x_i} = c_{1:D}(u_1, u_2, \cdots, u_D) \prod_{i=1}^{p} p_i(x_i) \tag{3}$$

where $c_{1:D}(u_1, u_2, \cdots, u_D)$ and $p_1(x_1), p_2(x_2), \cdots, p_D(x_D)$ denote the copula PDF and univariate marginal PDFs, respectively.

The conditional PDF can be obtained based on joint PDF through the operations of conditionalization and marginalization. Let $\boldsymbol{X}_{ta} \in \boldsymbol{\mathcal{R}}^{N_{ta}}$, $\boldsymbol{X}_o \in \boldsymbol{\mathcal{R}}^{N_o}$ and $\boldsymbol{X}_{uo} = \boldsymbol{X} - (\boldsymbol{X}_{ta} \cup \boldsymbol{X}_o) \in \boldsymbol{\mathcal{R}}^{D-N_{ta}-N_o}$ denote the vector corresponding to the target, observed and unobserved dimension of $\boldsymbol{X}$. In the case of complete information, $\boldsymbol{X}_{uo} = \phi$; and in the case of incomplete information, $\boldsymbol{X}_{uo} \neq \phi$. Given the available observation $\boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o$, the general solution of the predicted PDF after conditioning and marginalization operations is given by (Mu et al. 2020)

$$p(\boldsymbol{x}_{ta} | \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o) = \frac{p(\boldsymbol{x}_{ta}, \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o)}{p(\tilde{\boldsymbol{x}}_o)} = \frac{\int p(\boldsymbol{x}_{ta}, \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o, \boldsymbol{x}_{uo})\, d\boldsymbol{x}_{uo}}{\iint p(\boldsymbol{x}_{ta}, \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o, \boldsymbol{x}_{uo})\, d\boldsymbol{x}_{uo} d\boldsymbol{x}_{ta}} \tag{4}$$

In order to provides large solution space for PDF modelling of MUI geochemical data. Model class candidates of univariate marginal PDFs and copulas are constructed. The i-th candidate model of d-th univariate RV $\boldsymbol{X}_d$ is expressed as $\mathcal{M}_d^i (d = 1, \dots, D, i = 1, \dots, 13)$ associated with the corresponding parameter vector $\boldsymbol{\vartheta}_d^i (d = 1, \dots, D, i = 1, \dots, 13)$, where $\mathcal{M}_d^i$ are Normal kernel function ($\mathcal{M}_d^1$), Box kernel function ($\mathcal{M}_d^2$), Triangle kernel function ($\mathcal{M}_d^3$), Epanechnikov kernel function ($\mathcal{M}_d^4$), Normal distribution($\mathcal{M}_d^5$), Log-normal distribution($\mathcal{M}_d^6$), Weibull distribution($\mathcal{M}_d^7$), Gamma distribution($\mathcal{M}_d^8$), Gumbel distribution($\mathcal{M}_d^9$), Uniform distribution($\mathcal{M}_d^{10}$). This paper also introduces three kinds of Univariate Gaussian mixture models (UGMM) suitable for multimodal situations, namely: Bimodal UGMM($\mathcal{M}_d^{11}$), Tri-modal UGMM($\mathcal{M}_d^{12}$), Quad-modal UGMM($\mathcal{M}_d^{13}$). The m-th copula model candidate is expressed as $\mathcal{C}^m (m = 1,2)$ associated with the corresponding parameter vector $\boldsymbol{\psi}^m (m = 1,2)$, where $\mathcal{C}^m$ are multivariate Gaussian copula ($\mathcal{C}^1$) and multivariate Student's t copula ($\mathcal{C}^2$).

Bayesian inference on parameter level for $\boldsymbol{\vartheta}_d^i (d = 1, \dots, D, i = 1, \dots, 13)$ and $\boldsymbol{\psi}^m (m = 1,2)$ along with model level for $\mathcal{M}_d^i (d = 1, \dots, D, i = 1, \dots, 13)$ and $\mathcal{C}^m (m = 1,2)$ are conducted (Beck and Yuen 2004; Yuen 2010). Finally, the posterior multivariate joint CDF and PDF are given by:

$$F_{1:D}\big(x_1, x_2, \cdots, x_D | \widehat{\boldsymbol{\psi}}, \hat{\mathcal{C}}, \widehat{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\mathcal{M}}}\big) = C_{1:D}\big(\hat{u}_1, \hat{u}_2, \dots, \hat{u}_D | \widehat{\boldsymbol{\psi}}, \hat{\mathcal{C}}\big) \tag{5}$$

$$p_{1:D}\big(x_1, x_2, \cdots, x_D | \widehat{\boldsymbol{\psi}}, \hat{\mathcal{C}}, \widehat{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\mathcal{M}}}\big) = c_{1:D}\big(\hat{u}_1, \hat{u}_2, \dots, \hat{u}_D | \widehat{\boldsymbol{\psi}}, \hat{\mathcal{C}}\big) \cdot \prod_{d=1}^{D} p_d(x_d | \widehat{\boldsymbol{\vartheta}}_d, \widehat{\mathcal{M}}_d) \tag{6}$$

where $\widehat{\boldsymbol{\mathcal{M}}} = \big[\widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2, \dots, \widehat{\mathcal{M}}_D\big]$ is the collection of the most plausible univariate marginal PDFs, $\widehat{\boldsymbol{\vartheta}} = \big[\widehat{\boldsymbol{\vartheta}}_1, \widehat{\boldsymbol{\vartheta}}_2, \dots, \widehat{\boldsymbol{\vartheta}}_D\big]$ is the collection of the associated optimal parameter vectors of $\widehat{\boldsymbol{\mathcal{M}}}$, $\hat{\mathcal{C}}$ is the most plausible model of copula, and $\widehat{\boldsymbol{\psi}}$ is the associated optimal parameter vector of $\hat{\mathcal{C}}$. Details can be referred to (Mu et al. 2022).

2.2 Stage-CR

The conditional posterior PDF $p(\boldsymbol{x}_{ta} | \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o)$ can be obtained using Eq. (4). Note that $p(\boldsymbol{x}_{ta} | \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o)$ depends on $\widehat{\boldsymbol{\psi}}, \hat{\mathcal{C}}, \widehat{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\mathcal{M}}}$, but this is not reflected in notation due to expression simplicity. For the purpose of CR construction, the goal is to obtain a region of the sample given a cumulative probability α ($\alpha \in (0,100)$, unit: %). Due to the fact that this region is not unique when the joint PDF is asymmetric and/or multimodal, we introduce the idea of (Box and Tiao 2011) for CR construction that the goal is to find the region satisfying that any point inside this region should have probability density at least as large as any point outside the region. This is equivalent to the statement that this region occupies the smallest possible volume in the sample space. Let $R(p^\alpha)$ denote the subset of the sample space of $\boldsymbol{X}_{ta}$ of probability density p^α:

$$Re(p^\alpha) = \{\boldsymbol{x}_{ta} : \boldsymbol{x}_{ta} \in \mathbb{R}^{N_{ta}}, p(\boldsymbol{x}_{ta} | \boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o) \ge p^\alpha\} \tag{7}$$

Consider the following arguments of the maxima:

$$\hat{p}^{\alpha} = \arg\max_{p^{\alpha}} \int_{Re(p^{\alpha})} p(\boldsymbol{x}_{ta}|\boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o) d\boldsymbol{x}_{ta} = 1 - \alpha \tag{8}$$

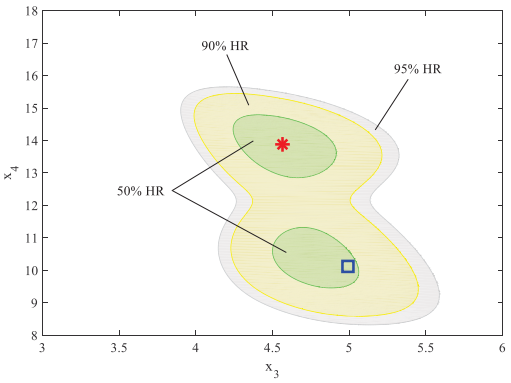
Then, $Re(\hat{p}^{\alpha})$ is the α ($\alpha \in (0,100)$, unit: %) Highest posterior density Regions (HR), and the boundary of $Re(\hat{p}^{\alpha})$ is the $100(1-\alpha)$% Highest posterior density Contour (HC), denoted as $Co(\hat{p}^{\alpha})$. There is no general analytic solution for the above optimization, so numerical integration (Haselsteiner et al. 2017; Wright 1986) or Monte Carlo Simulation (MCS) (Hyndman 1996) is required. Here, MCS samples of $p(\boldsymbol{x}_{ta}|\boldsymbol{X}_o = \tilde{\boldsymbol{x}}_o)$ are simulated for obtaining the HR $Re(\hat{p}^{\alpha})$ and $Co(\hat{p}^{\alpha})$.

3 Illustrated examples

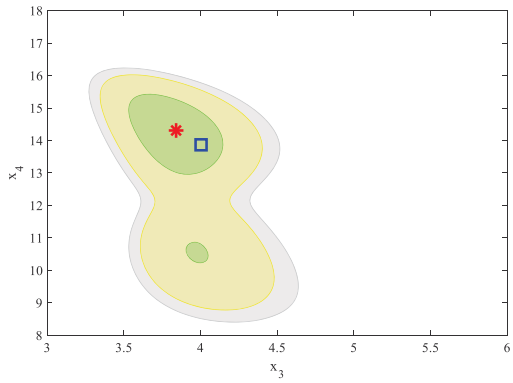
This simulated example utilizes the BASIC-HR for a four-dimensional correlated RVs $\boldsymbol{X} = [X_1, \dots, X_4]^T$ with the joint PDF of asymmetricity and multimodality. Consider the following linear transformation:

$$\boldsymbol{X} = \begin{bmatrix} 0.6 & 0 & 0.2 & 0.2 \\ 0 & 0.7 & 0.3 & 0 \\ 0.2 & 0.3 & 0.5 & 0 \\ 0.2 & 0 & 0 & 0.8 \end{bmatrix} \boldsymbol{Z} \tag{9}$$

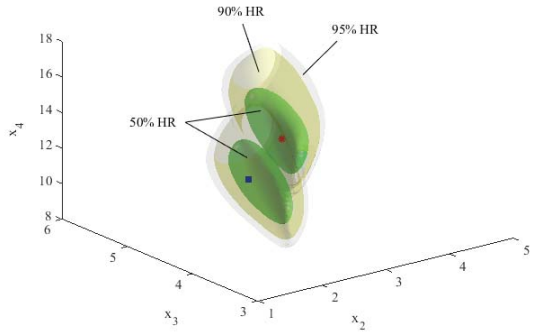
where $\boldsymbol{Z} = [Z_1, \dots, Z_4]^T$ is a RV with four uncorrelated components. The PDFs of $Z_1, \dots, Z_4$, are Normal, Uniform, Gamma and Bimodal UGMM distributions, respectively. The purpose of including various types of distributions for $\boldsymbol{Z}$ is to mimic the complex statistical behaviors of MUI geotechnical data.



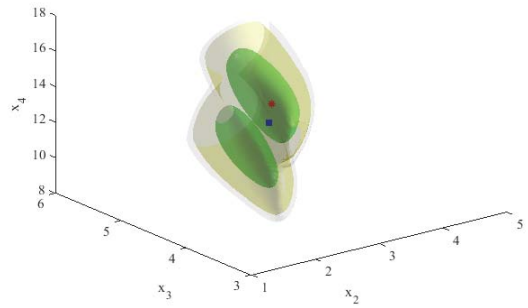
(c) $p(x_3, x_4|X_1 = \tilde{x}_1, X_2 = \tilde{x}_2)$ of 2-nd data point



(d) $p(x_3, x_4|X_1 = \tilde{x}_1, X_2 = \tilde{x}_2)$ of 192-nd data point



(e) $p(x_2, x_3, x_4|X_1 = \tilde{x}_1)$ of 2-nd data point



(f) $p(x_2, x_3, x_4|X_1 = \tilde{x}_1)$of 192-nd data point

Figure 1. Predicted PDFs and 50% HR, 90% HR, 95% given different information of different data points.

Figure 1 shows predicted PDFs and 50% HR, 90% HR, 95% given different information of different data points. The true value and optimal value are denoted by blue square and red asterisk, respectively. The 50% HR, 90% HR, and 95% HR are shown by green area, yellow area and gray area, respectively. Although $\boldsymbol{X}$ is just the linear transformation of $\boldsymbol{Z}$ with each component $Z_i, i = 1, \dots, 4$ being a common probability distribution, the joint posterior PDF of $\boldsymbol{Z}$ as well as the conditional posterior PDF are very irregular. The asymmetricity inherits from Z_2 and Z_3 while the bimodality inherits from Z_4. The true value is within 50% HR, confirming the modelling and prediction capabilities of the proposed BASIC-HR. It can be observed that 50% HR is composed by two separated regions. This is reasonable because the bimodality of the PDF implying that the high posterior density regions

concentrate at two peaks. It is worth noting that for 50% HRs of two left subplots, the true values do not locate in the same region of the optimal value. Here, if a traditional CR, for example, a CR defined by the optimal value plus and minus a magnification factor times the standard deviation, is adopted, the prediction region will around the region of the optimal value, highly possibly leading to the consequence that the adopted traditional CR does not include the true value. This unexpected consequence will cause unreliable design. In conclusion, the proposed BASIC-HR is particularly suitable for a dataset with MUI characteristics.

4 Conclusion

This paper proposes a unified framework (called BASIC-HR) for PDF modelling and CR construction based on simulated geotechnical data with MUI characteristics. Stage-PDF obtains the posterior joint CDF and PDF by fusing copula theory and Bayesian inference. Stage-CR introduces the HR, which is defined as that, given a cumulative probability, every point inside the HR should have posterior probability density at least as large as every point outside the HR. The illustrated example of simulating data with MUI characteristics shows that BASIC-HR is capable of properly modelling the asymmetric and multimodal PDF and constructing rational CRs for reliable design.

Acknowledgments

This work was supported by Guangdong Provincial Key Laboratory of Modern Civil Engineering Technology (2021B1212040003) and Guangdong Hong Kong-Macau Joint Laboratory Program (Project No.: 2020B1212030009). This generous support is gratefully acknowledged.

References

Beck, J. L., and Yuen, K. V. (2004). "Bayesian Probabilistic Approach." *Journal of Engineering Mechanics*, 130. No. 2(February), 192–203.

Box, G. E. P., and Tiao, G. C. (2011). Bayesian inference in statistical analysis. *John Wiley & Sons*.

Ching, J., and Phoon, K.-K. (2020). "Constructing a Site-Specific Multivariate Probability Distribution Using Sparse, Incomplete, and Spatially Variable (MUSIC-X) Data." *Journal of Engineering Mechanics*, 146(7), 04020061.

Haselsteiner, A. F., Ohlendorf, J.-H., Wosniok, W., and Thoben, K.-D. (2017). "Deriving environmental contours from highest density regions." *Coastal Engineering*, Elsevier, 123, 42–51.

Hyndman, R. J. (1996). "Computing and graphing highest density regions." *The American Statistician, Taylor & Francis*, 50(2), 120–126.

Mu, H.-Q., Shen, J.-H., Zhao, Z.-T., Liu, H.-T., and Yuen, K.-V. (2022). "A novel generative approach for modal frequency probabilistic prediction under varying environmental condition using incomplete information." *Engineering Structures*, *Elsevier*, 252, 113571.

Mu, H. Q., Liu, H. T., and Shen, J. H. (2020). "Copula-based uncertainty quantification (Copula-uq) for multi-sensor data in structural health monitoring." *Sensors (Switzerland)*, 20(19), 1–18.

Nelsen, R. B. (2007). An introduction to copulas. *Springer Science & Business Media*.

Sklar, A. (1996). "Random variables, distribution functions, and copulas: a personal look backward and forward." *Lecture notes-monograph series, JSTOR*, 1–14.

Sklar, M. (1959). "Fonctions de repartition an dimensions et leurs marges." *Publ. inst. statist. univ. Paris*, 8, 229–231.

Wang, Y., Cao, Z., and Li, D. (2016). "Bayesian perspective on geotechnical variability and site characterization." *Engineering Geology, Elsevier B.V.*, 203, 117–125.

Wang, Y., Huang, K., and Cao, Z. (2013). "Probabilistic identification of underground soil stratification using cone penetration tests." *Canadian Geotechnical Journal*, 50(7), 766–776.

Wright, D. E. (1986). "A note on the construction of highest posterior density intervals." *Journal of the Royal Statistical Society: Series C (Applied Statistics), Wiley Online Library*, 35(1), 49–53.

Yuen, K.-V. (2010). Bayesian methods for structural dynamics and civil engineering. *John Wiley & Sons*.