



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# **On the use of random sets in geotechnical engineering**

**Juan José Sepúlveda García**

Universidad Nacional de Colombia  
Faculty of Engineering  
Department of Civil and Agricultural Engineering  
Bogotá D.C., Colombia  
2021



# **On the use of random sets in geotechnical engineering**

**Juan José Sepúlveda García**

jjsepulvedag@unal.edu.co

Dissertation submitted in partial fulfillment of the requirements for the degree of:

**Master in Engineering - Geotechnics**

Advisor:

**Dr. Techn. Diego Andrés Álvarez Marín**

daalvarez@unal.edu.co

Research group:

Earthquake Engineering and Seismology Research Group, UN Manizales

Universidad Nacional de Colombia

Faculty of Engineering

Department of Civil and Agricultural Engineering

Bogotá D.C., Colombia

2021



With love, to my family ♥



# Acknowledgements

First of all, I would like to express my eternal gratitude to my advisor, Ph.D. Diego Andrés Álvarez Marín, for his patience, understanding, commitment, and countless advice in the development of this project. I am totally sure that without his help and support this work would never have developed. Thank you very much for trusting me, for encouraging me, and for accompanying me throughout this process. His support has been invaluable, and his mentoring has taught me a lot not only academically but also personally.

Second, I would like to express my infinite gratitude and love to my family. Their unconditional love and support during every stage of my life have been fundamental in my development. Additionally, without their collaboration and help during these last two years, the completion of this degree work would have been impossible. I know that I can always count on you, and that you are there for me in my best and worst moments, for that I am enormously grateful to you. There may be differences and problems, but family is family and I thank God he gave me the best in the world. You cannot imagine how much I love you.

Third, I would like to thank a group of friends in whom I always found support and encouraging words in the most difficult moments of this process. Thanks to Angélica Betancourth Arias, Alejandra Mendoza Rojas, Omar Andrés Rosada González, Daniela Alejandra Morales Herrera, Yesenia Chacón Gordillo, and Daniela Arias López. I am extremely grateful for your unconditional friendship, for your good wishes, for your encouraging words, and because you endured more than one of my many existential crises. You will always be my friends, and even if we take different directions in life, I will always be grateful to you and you can count on me. I owe you all some good beers.

Last but not least, I express my eternal gratitude and admiration for the Universidad Nacional de Colombia. This glorious institution trained me as a professional and I have the fondest memories of my university life from it. I feel privileged to be a graduate of such an illustrious institution and, for the same reason, I feel a great responsibility that I will assume with gallantry. I hope that for the rest of my life I can give back to society everything that the Universidad Nacional de Colombia gave me.



# Abstract

Geotechnics is subject to two major types of uncertainty: aleatory and epistemic. Aleatory uncertainty refers to the inherent variability of materials and their external agents, while epistemic uncertainty refers to the lack of information and knowledge. A comprehensive geotechnical model should take into account these two types of uncertainties in light of a reliability analysis. However, the reliability analyses that are conventionally performed in geotechnical engineering have a series of gaps and approximations that may diverge model estimations from reality. In the first place, reliability analyses have been commonly based on probability theory, which is capable of modeling aleatory but not epistemic uncertainty, so the latter is usually ignored. Second, assumptions are made about the dependence between the basic variables of the models, which are not validated and in most cases are unfounded. Third, obtaining accurate modeling results requires excessive computational costs, which in many cases are unfeasible.

In order to fill these gaps in the *state-of-the-art*, this thesis proposes a methodology for the evaluation of reliability in geotechnics, which is computationally efficient, takes into account the dependence between the basic variables and also their aleatory and epistemic uncertainty. Specifically, subset simulation is employed for efficient and accurate calculation of probabilities of failure, copula theory is used to model dependence among basic variables, and random set theory is employed to model both epistemic and aleatory uncertainties.

The proposed methodology manages to integrate the three aforementioned developments in a single approach for the analysis of the reliability in geotechnical models. The applicability of this procedure is proved through different practical examples of geotechnical engineering. The results show the efficiency of the proposed algorithm, the importance of dependence in reliability analyses, and the impact that aleatory and epistemic uncertainties have on the final results of the modeling. In conclusion, the proposed method proves to be a fairly complete tool with a very wide application range, so that geotechnical engineers will have the possibility of implementing it in their designs and modeling (in fact, they should).

**Keywords:** uncertainty; reliability analysis; monte carlo simulation; subset simulation; copulas; random sets theory; imprecise probabilities; lower and upper probabilities of failure.

# Resumen

La geotecnia está sujeta a dos grandes tipos de incertidumbre: aleatorias y epistémicas. La incertidumbre aleatoria se manifiesta a través de la variabilidad inherente de los materiales y de sus agentes externos, mientras que la incertidumbre epistémica se manifiesta a través de la escasez de la información y de la falta de conocimiento. Un modelo geotécnico integral debería de tener en cuenta estos dos tipos de incertidumbre a la luz de un análisis de confiabilidad. No obstante, los análisis de confiabilidad que se desarrollan convencionalmente en la ingeniería geotécnica tienen una serie de vacíos y aproximaciones que pueden generar que los resultados de los modelos disten de la realidad. En primer lugar, los análisis de confiabilidad se han basado comúnmente en la teoría de la probabilidad, la cual es capaz de modelar la incertidumbre aleatoria pero no la epistémica, por lo que esta última se suele obviar. En segundo lugar, se realizan supuestos sobre la dependencia entre las variables básicas de los modelos, las cuales no son validadas y en la mayoría de los casos carecen de fundamentos. En tercer lugar, obtener resultados precisos requiere de un costo computacional excesivo, que en muchos casos puede ser inviable.

Con el objetivo de suplir estos vacíos en el estado del arte, esta tesis propone una metodología para la evaluación de la confiabilidad en geotecnia, la cual es eficiente computacionalmente, tiene en cuenta la dependencia entre las variables básicas y también su incertidumbre aleatoria y epistémica. Específicamente, se usa el algoritmo subset simulation para el cálculo eficiente y preciso de las probabilidades de falla, se hace uso de la teoría de copulas para modelar la dependencia entre las variables, y se emplea la teoría de random sets para modelar la incertidumbre epistémica y aleatoria.

La metodología propuesta logra integrar los tres desarrollos anteriormente mencionados en un único enfoque para el análisis de la confiabilidad de modelos geotécnicos. La aplicabilidad de este enfoque se demuestra a través de diferentes ejemplos prácticos de la ingeniería geotécnica. Los resultados evidencian la eficiencia del algoritmo propuesto, la importancia de la dependencia en los análisis de confiabilidad, y el impacto que las incertidumbres aleatorias y epistémicas tienen en las modelaciones. En conclusión, el enfoque propuesto es una herramienta bastante completa y con una aplicación muy amplia para realizar análisis de confiabilidad, por lo que los geotecnistas tendrán la posibilidad de implementarla en sus diseños y modelaciones (de hecho, ellos deberían usarla).

**Palabras clave:** incertidumbre, análisis de confiabilidad, simulación monte carlo, simulación de subconjuntos, copulas, teoría de conjuntos aleatorios, probabilidades imprecisas, probabilidades de falla superior e inferior.

# Contents

<b>Acknowledgements</b>	<b>vii</b>
<b>Abstract</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem statement . . . . .	6
1.3 Aim and objectives . . . . .	7
1.4 Document outline . . . . .	7
<b>2 Fundamentals of probability and reliability</b>	<b>8</b>
2.1 Some concepts of probability theory . . . . .	8
2.1.1 Axioms . . . . .	8
2.1.2 Random variables . . . . .	9
2.1.3 Distribution functions . . . . .	10
2.1.4 Joint distribution functions . . . . .	12
2.1.5 Correlation among random variables . . . . .	14
2.1.6 Autocorrelation . . . . .	14
2.1.7 Markov Chains . . . . .	15
2.2 Introduction to reliability analysis . . . . .	19
2.2.1 The safety challenge . . . . .	19
2.2.1.1 Deterministic analysis . . . . .	20
2.2.1.2 Semi-probabilistic analysis . . . . .	21
2.2.1.3 Probabilistic analysis . . . . .	22
2.2.2 Computing probability of failure . . . . .	25
2.2.2.1 Numerical integration . . . . .	25
2.2.2.2 Transformation methods . . . . .	26
2.2.2.3 Simulation methods . . . . .	27
<b>3 Subset simulation for assessing geotechnical reliability</b>	<b>30</b>
3.1 Basic idea of subset simulation . . . . .	30
3.2 Markov-Chain Monte Carlo methods . . . . .	32
3.2.1 Metropolis algorithm . . . . .	33

3.2.2	Metropolis-Hastings algorithm . . . . .	35
3.2.3	Techniques to improve Markov Chains . . . . .	37
3.2.3.1	Chain starting value . . . . .	37
3.2.3.2	Definition of the proposal distribution . . . . .	37
3.2.3.3	Length of the chain . . . . .	38
3.2.3.4	Burn-in period . . . . .	38
3.2.3.5	Thinning or lag period . . . . .	38
3.2.3.6	Test for convergence . . . . .	39
3.3	Subset simulation . . . . .	42
3.3.1	Choice of intermediate failure events . . . . .	43
3.3.2	Modified Metropolis algorithm . . . . .	44
3.3.3	Estimation of the probability of failure . . . . .	47
3.4	Practical examples of subset simulation . . . . .	48
3.4.1	Parabolic performance function . . . . .	49
3.4.2	Shallow footing with variable width . . . . .	51

<b>4</b>	<b>On the dependence and the use of copulas in geotechnics: a state-of-the-art review</b>	<b>55</b>
4.1	Contextualization . . . . .	55
4.2	Introduction to copula theory . . . . .	58
4.2.1	Preliminary concepts . . . . .	58
4.2.1.1	Increasing functions . . . . .	58
4.2.1.2	Quasi-inverses . . . . .	58
4.2.1.3	$V_H$ -volume . . . . .	58
4.2.1.4	Multivariate cumulative distribution functions . . . . .	59
4.2.1.5	$k$ -margins . . . . .	59
4.2.2	Copula functions . . . . .	60
4.2.3	Sklar's theorem . . . . .	61
4.2.4	The Frechét-Hoeffding bounds . . . . .	62
4.2.5	Some copula properties . . . . .	63
4.2.5.1	Range of dependence values . . . . .	63
4.2.5.2	Radial symmetry . . . . .	64
4.2.5.3	Tail dependence . . . . .	65
4.2.6	Empirical copulas . . . . .	65
4.2.7	Vine copulas . . . . .	66
4.3	Some remarks about dependence in geotechnics . . . . .	70
4.3.1	The Nataf transformation . . . . .	72
4.4	Construction of copulas for modeling dependence among geotechnical variables . . . . .	74
4.4.1	Selection of marginal distributions . . . . .	75

---

4.4.2	Measures of correlation . . . . .	76
4.4.2.1	Linear correlation . . . . .	77
4.4.2.2	Concordance . . . . .	78
4.4.3	Popular families of copulas . . . . .	80
4.4.3.1	Elliptical copulas . . . . .	80
4.4.3.2	Archimedean copulas . . . . .	83
4.4.3.3	Plackett copula . . . . .	87
4.4.4	Estimation of the parameters of a copula . . . . .	88
4.4.4.1	Method of moments based on rank correlations . . . . .	88
4.4.4.2	Maximum Likelihood Estimator . . . . .	90
4.4.4.3	The inference functions from margins and the canonical maximum likelihood methods . . . . .	92
4.4.5	Tests of goodness-of-fit for copulas . . . . .	93
4.4.5.1	Goodness-of-fit test based on Empirical Copula . . . . .	95
4.4.5.2	Akaike and Bayesian information criterion . . . . .	96
4.5	Some applications of copula theory in geotechnical engineering . . . . .	97
4.5.1	Shear strength parameters of the Mohr-Coulomb failure criterion . . . . .	97
4.5.2	Settlement . . . . .	106
4.5.3	Geotechnical earthquake parameters . . . . .	115
4.5.4	Soil spatial modelling . . . . .	117
4.5.5	Other uses of copulas in geotechnics . . . . .	120
4.5.5.1	Landslide hazard assessments . . . . .	120
4.5.5.2	Earthquake early warning . . . . .	121
4.5.5.3	Studies on sample size . . . . .	122
4.6	Uncertainties in the construction and definition of copula for geotechnical variables	122
4.7	Reliability analysis in geotechnics considering dependence through copulas . . . . .	128
4.7.1	The copula based sampling and copula direct integration methods. . . . .	129
4.7.1.1	Direct integration method in copulas . . . . .	129
4.7.1.2	Monte Carlo Simulation in copulas . . . . .	131
4.7.2	Results of reliability analysis . . . . .	132
4.7.3	Efforts to unify results of reliability analysis of geotechnical structures when dependence is considered. . . . .	138
<b>5</b>	<b>Random sets and evidence theory in geotechnical engineering</b>	<b>144</b>
5.1	Dempster-Shafer evidence theory . . . . .	145
5.1.1	Belief and Plausibility measures . . . . .	147
5.2	Succinct introduction to random sets theory . . . . .	149
5.2.1	From random variables to random sets . . . . .	149

5.2.2	Some representations of the state of knowledge and their relationship with random sets theory . . . . .	150
5.2.2.1	Possibility distributions . . . . .	151
5.2.2.2	Probability boxes . . . . .	152
5.2.2.3	Families of intervals . . . . .	155
5.2.2.4	Probability functions . . . . .	155
5.2.3	Combination of evidence . . . . .	156
5.2.4	Random relations for random sets . . . . .	156
5.2.5	Dependence considerations . . . . .	157
5.2.6	Extension principle for random sets . . . . .	158
5.2.6.1	Method of optimization . . . . .	159
5.2.6.2	Method of sampling . . . . .	159
5.2.6.3	Method of vertices . . . . .	160
5.2.6.4	Method of function approximation . . . . .	160
5.3	Employability of random sets theory in geotechnical engineering . . . . .	160
5.3.1	Practical example on the use of random sets in a geotechnical model . . . . .	161
<b>6</b>	<b>Proposed approach for rigorous and efficient geotechnical reliability analyses</b>	<b>171</b>
6.1	Theoretical basis of the procedure . . . . .	171
6.1.1	Indexation and sampling from a random sets . . . . .	172
6.1.1.1	Indexation and sampling from a possibility distribution . . . . .	174
6.1.1.2	Indexation and sampling from a distribution-free probability box . . . . .	174
6.1.1.3	Indexation and sampling from a CDF . . . . .	175
6.1.1.4	Indexation and sampling from a finite family of intervals . . . . .	175
6.1.2	Bounding of the probability of failure using random sets . . . . .	176
6.1.3	Monte Carlo simulation for the estimation of $P_f$ and $\overline{P}_f$ . . . . .	177
6.2	Proposed approach: efficient estimation of $P_f$ and $\overline{P}_f$ using subset simulation . . . . .	179
6.3	Numerical examples . . . . .	180
6.3.1	Example 1: Rotational failure of an homogeneous soil slope . . . . .	180
6.3.2	Example 2: Toppling failure of a rock slope . . . . .	183
6.3.3	Example 3: Plane failure of a rock slope . . . . .	190
<b>7</b>	<b>Conclusions and final considerations</b>	<b>195</b>
7.1	Comments on the use of subset simulation in geotechnics . . . . .	197
7.2	Comments on the dependence and copula theory in geotechnics . . . . .	198
7.3	Comments on the applicability of random sets theory in geotechnical engineering . . . . .	200
7.4	Conclusions on the proposed methodology for reliability analyses in geotechnics . . . . .	202
7.5	Open issues and directions for future research . . . . .	203
	<b>Bibliography</b>	<b>221</b>

# 1 Introduction

## 1.1 Motivation

Every civil construction and intervention is related, to a greater or lesser extent, with soils or rocks. Some structures are supported on these materials (e.g. buildings or bridges), some others support them (e.g. retaining walls or tunnels), and even some others are made up with them (e.g. embankments and slopes). This fact is indicative of the importance of soils and rocks in the current civil engineering practice, since the stability and correct performance of the different structures is conditioned by the understanding of these natural materials ([Lambe and Whitman, 1991](#); [Mitchell et al., 2005](#)).

However, unlike man-made materials, whose elaboration is outlined and hence their properties are known within narrow limits, engineers do not control the genesis of soils and rocks ([Mitchell et al., 2005](#)). These natural materials result from a great variety of physical and chemical processes and through their history have been subjected to several internal and external agents that condition their behavior (see e.g. [Blyth and De Freitas, 2017](#); [Tarbuck et al., 2005](#)). Consequently, soils and rocks exhibit significant *randomness* in their properties and loads, and their behavior has an implicit degree of *uncertainty*, even for the same type of material or deposit.

To understand the implications of the inherent nature of these materials, it is important to define what the terms *random* and *uncertain* mean. In the Merriam-Webster dictionary both words are defined as

**Random** (adjective). Date: 1565. 1. a: lacking a definite plan, purpose, or pattern b: made, done, or chosen at random; 2. a: relating to, having, or being elements or events with definite probability of occurrence. b: being or relating to a set or to an element of a set each of whose elements have equal probability of occurrence. ([Staff, 2004](#))

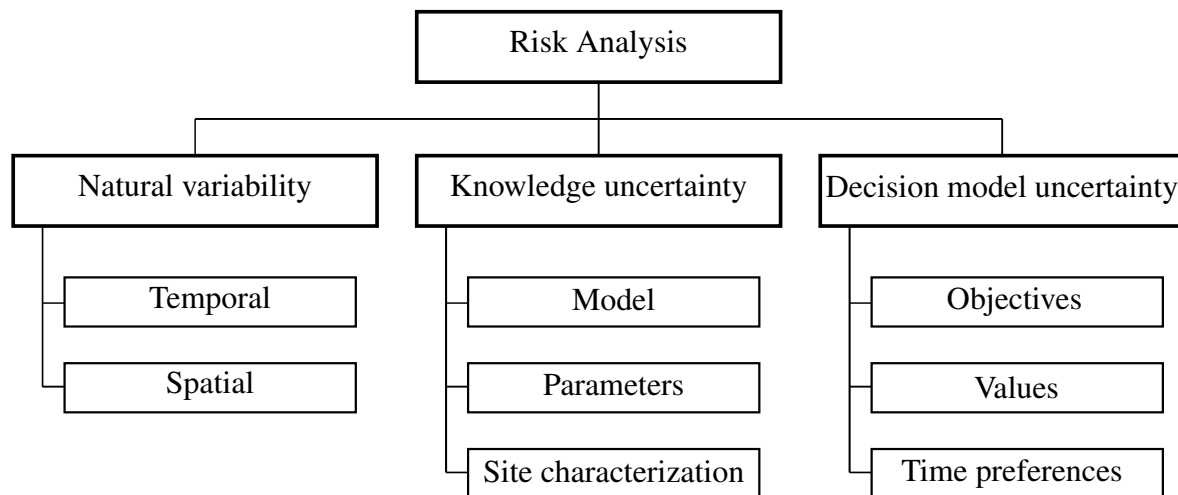
**Uncertain** (adjective). Date: 14th century. 1: Indefinite, indeterminate 2: not certain to occur: Problematical 3: not reliable: Untrustworthy 4: a: not known beyond doubt: Doubtful b: not having certain knowledge: Doubtful c: not clearly identified or defined 5: not constant: Variable, Fitful. ([Staff, 2004](#))

In this sense, randomness is associated with a lack of patterns or with unpredictable events caused by *change*, states that are innate to nature and independent of anyone's knowledge. Meanwhile,

uncertainty refers to the lack of total certainty, or to a limited knowledge, which makes it impossible to exactly describe a current state or a future outcome, even though it is possible to predict it (Baecher and Christian, 2005).

Helton (1997) classified uncertainty in two major categories, namely *aleatory* and *epistemic*. On the one hand, aleatory uncertainty (also known as stochastic uncertainty, irreducible uncertainty, or variability) refers to the natural variability or randomness of a system and its variables. On the other hand, epistemic uncertainty (also known as subjective uncertainty, reducible uncertainty, state of knowledge uncertainty, or ignorance) alludes to a lack of knowledge about data or a system. Note that while aleatory uncertainty is typical of materials and loads that make up the system, and therefore it can not be reduced, epistemic uncertainty is inherent to the available knowledge, and therefore it can be reduced by gathering more and better information. In this way, these two concepts encompass what has been previously defined as randomness and uncertainty, and hence, the two categories defined by Helton (1997) will also be adopted hereinafter in this document.

Particularly in the context of geotechnics, Baecher and Christian (2005) alluded to these two main categories of uncertainty and disaggregated them into their respective components according to risk analyses conducted in geotechnical engineering. Figure 1-1 shows a concept map that breaks down the main categories of uncertainty in geotechnics.



**Figure 1-1:** Categories of uncertainties in geotechnical risk analysis (adapted from Baecher and Christian, 2005)

According to Baecher and Christian (2005), uncertainty in risk analysis is divided into natural variability (i.e. aleatory uncertainty), knowledge uncertainty (i.e. epistemic uncertainty), and decision model uncertainty. Natural variability refers to the inherent randomness and is made up of (1) temporal uncertainty and (2) spatial uncertainty, in the sense that soils and rocks are variable over time (at a single location) and over space (at a single time). Knowledge uncertainty refers to a lack of data and/or knowledge that is reflected in: (1) model uncertainty since models are just

approximations that mimic reality, may vary over time, and sometimes the best model is difficult to define, (2) parameter uncertainty since the parameter's estimates depends on both the quality and quantity of information, and (3) site characterization uncertainty that alludes to the suitability of the interpretations about the surface geology, and results from explorations uncertainties such as measurement errors, the inconsistency of data, data handling and transcription errors and inadequate samples that may not adequately represent the soil/rock. Finally, decision model uncertainty has to do with social, economic, and operational issues, and hence it will be obviated hereinafter in this document since it does not concern the focus of this research.

In this sense, when constructing geologic-geotechnical models, analysts are faced with these two main types of uncertainties, which represent a formidable challenge for guaranteeing the correct performance of the geotechnical systems. Unfortunately, civil engineering has traditionally been taught and applied from a very deterministic point of view, as a precise science that can describe any current state and anticipate any future outcome. However, engineering is not a science but rather an art and, under randomness and lack of knowledge, it only leads to vague approximations of reality (Phoon, 2020).

This is how it is necessary to complement the purely deterministic analyses with concepts such as the *factor of safety (FS)*, that seeks to obtain safer structures and systems. In its most general form, the factor of safety is defined as the ratio of the applied load over a system to its resisting forces. Thus, from a purely theoretical perspective, a geotechnical model will be stable if  $FS > 1$ , will fail if  $FS < 1$ , and will be in limit equilibrium if  $FS = 1$ . However, as in practice there are implicit uncertainties in the deterministic designs, codes and norms require the minimum factors of safety to be greater than 1, and whose value depends on the nature of the system, the risks involved, and the type of analysis conducted. The magnitude of the minimum factor of safety stems from past experiences in similar systems, but in turn, they are conditioned by economical aspects, social factors, and particular conditions of each region or country.

Despite that the use of factors of safety is widely extended, and even some countries require their mandatory use in practice, these factors have several disadvantages, many of them related to the handling of uncertainty. A factor of safety does not give information about uncertainties, their sources, where they are concentrated, and how to handle or reduce them. So, when things go wrong in a purely deterministic design, this may take the analyst by surprise. Thus, in some situations where the uncertainties are many, a very large factor of safety must be employed, which may be too conservative and leads to expensive constructions. Due to this, other complementary methods have appeared in geotechnics, such as the popular *observational method* (see Peck, 1969; Terzaghi et al., 1996).

The observational method consists of continuous monitoring before, during, and even after geotechnical constructions or interventions. In this continuous monitoring, some crucial parameters in designs must be measured, such as settlements, water pressures, presence of cracks or weak layers, etc. If some of these parameters deviate from the conditions initially stated in the original design,

this design must be revised and modified as new information appears. Thus, the observational method avoids the use of too large factors of safety, since the original model is getting refined as new real-time information is obtained. Although this method has many advantages and should be employed whenever possible, it also has some drawbacks. Some of its disadvantages are the extension of the intervention times, inasmuch as the designs may be modified and the construction reevaluated several times, or that some modifications, in light of new information, are unfeasible contractually (see [Peck, 1969](#), for more details).

None of the two forementioned methods manage to quantify and propagate uncertainties in the geotechnical models. On the one hand, fully deterministic analyses implicitly handle uncertainty with a factor of safety that may be quite large. On the other hand, the observational method gives better handling to uncertainty through continuous monitoring, but again, uncertainty is not quantified nor propagated in models at any moment. In this way, if the engineers want to assess uncertainties and study their implications in the models and design, they have to use other methodologies. In particular, reliability analyses must be carried out, most of which are based on probability theory.

A reliability assessment, based on probability theory, aims to estimate the probability of failure  $P_f$ , that is, the probability of an undesirable performance of a system ([Melchers and Beck, 2018](#); [Ditlevsen and Madsen, 1996](#)). Unlike deterministic methods, a probabilistic analysis models the variables involved in the systems as random variables instead of constant values. In this way, analysts can explicitly handle uncertainty, propagate it through the system, and evaluate its implications in the final designs. However, despite this possibility and its benefits, engineers have been reluctant to apply probabilistic analysis in geotechnical practice for two main reasons: (1) it involves some concepts and terminology which are not familiar to many engineers, and (2) it is wrongly supposed that this type of assessments requires more data, time and effort ([Phoon and Ching, 2014](#), chap. 3). Nonetheless, none of these arguments is valid to avoid explicit handling of uncertainty.

Probabilistic analysis does not add and significant complexity to models, neither needs more data nor time nor effort in contrast to the commonly employed deterministic analysis; therefore, it is a tool which all geotechnical engineers should be aware of. However, reliability assessments based on purely probabilistic analysis have some disadvantages or obstacles in its application. Among the main difficulties in probabilistic analysis for purposes of reliability assessments, there are:

1. After identifying the random variables involved in the system and constructing their multivariate probability function, it is regularly found that the failure domain is on the tails of this distribution. That is to say, the respective  $P_f$  is associated with rare or infrequent events represented by the tails of the probability distribution. Commonly employed simulation methods, such as Monte Carlos simulation (MCS), hardly manage to sample the tails of distributions, so in order to obtain relative accurate approximations of  $P_f$  too large sample sizes must be employed. Consequently, computation times may be extremely large and even

so the  $P_f$  may be not accurately defined.

2. Although defining univariate probability models for the random variables is relatively straightforward, this is not the case of multivariate probability models, i.e. models that relate all the random variables involved in the system. Independence among random variables is frequently assumed in order to simplify calculations, but this assumption is neither justified nor corroborated. In other cases, the multivariate normal distribution is assumed, again neither with justification nor validation, to represent the relation of all the random variables. These two assumptions are quite frequent in probabilistic analysis, but both are too restrictive and may lead to failure probabilities that do not represent reality.
3. Probability theory has proven to be a good tool to model aleatory uncertainty but not epistemic uncertainty. Sometimes, efforts are made to express epistemic uncertainty in terms of aleatory uncertainty in order to propagate it through the models, but these efforts are not fruitful at all, as will be explained later. In light of scarce information, as it is common in geotechnics, epistemic uncertainty may play a role of equal or higher importance than aleatory uncertainty. Therefore, a framework is needed to represent and propagate epistemic uncertainty in the models, in conjunction with aleatory uncertainty.

A comprehensive reliability analysis has to overcome these obstacles, so it must include epistemic and aleatory uncertainty, have a wide range of multivariate models from which to choose from and be computationally efficient when computing the failure probability. Thankfully, in the last decades there have been developments of some methodologies that can face these issues. Particularly, some of these developments are suitable to be applied in practice in order to complement standard probability analysis.

In the case of an efficient and accurate way to compute probabilities of failure given a certain model, Subset simulation (SubSim) has proven to be an appropriate methodology for this end. Subset simulation is a methodology that follows the nature of the MCS, but unlike this one, it is especially appropriate to sample the tails of the distributions, where the domain of failure is found. This methodology is of general applicability just like MCS, so it can be employed in geotechnical engineering to compute the probability of failure of any model. In general, SubSim has the same benefits and applicability of MCS, but it is considerably much less computationally expensive.

On the other hand, copula theory has demonstrated to be the theory of choice for constructing multivariate probability models. Copula theory splits multivariate probability distributions into its marginal distributions (univariate) and a structure of dependence (the so-called copula). Thus, the problem of defining a multivariate model is broken down into the selection of suitable univariate distributions for each random variable and the selection of a copula that groups all variables. There exists a lot of copula functions in the literature, each one with its peculiarities, advantages, and disadvantages when representing a certain group of random variables. So, this theory offers a great variety of multivariate probability models, and of course encompass the common assumptions of independence or gaussian dependence.

Finally, there are plenty amount of theories that aim to treat epistemic uncertainty, but among all of them the random set theory excels as a well suited theory to jointly handle epistemic and aleatory uncertainty. This theory is a generalization of the conventional probability theory, where a realization of a random variable is a subset of a given domain, instead of a single value. This structure manages to include both aleatory and epistemic uncertainty in the representation of a random variable. Additionally, random set theory is also able of handling different types and structures of information such as intervals, possibility distributions, probability boxes, Dempster Shafter structures and probability density functions. Thus, the geotechnist will be able to take into account different sources of information such as laboratory or field tests, data from other nearby projects, and even the opinion of local experts expressed as intervals or in linguistic terms.

It is worth noting that the above three methodologies are independent, and hence any of these could be applied in solitary to a reliability assessment. However, by implementing in conjunction the above three methodologies, the geotechnical reliability assessments will be comprehensive, and it will overcome the limitations of the conventional probabilistic analyses.

## 1.2 Problem statement

Currently, geotechnical engineers aware of uncertainties in their designs and constructions must appeal to reliability assessments, conducted through a probability approach, in order to propagate uncertainty and to estimate the probability of failure. However, probabilistic methods have some drawbacks and assumptions that may lead to a reliability assessment far from reality, either with conservative or unconservative estimates.

In the last decades, there have been some developments that allow analysts to overcome some of these pitfalls and assumptions. Particularly, SubSim overcomes the main issues of a MCS by efficiently sampling the tails of the distributions and accurately estimating probabilities of failure. Copula theory offers a new framework for constructing multivariate probability distributions, where the commonly applied assumptions of independence or gaussian dependence are just one option of the wide spectrum of possibilities. Finally, random set theory allows to include epistemic uncertainty in the estimates of probabilities of failure, thus overcoming one of the major disadvantages of reliability assessments conducted through a purely probabilistic approach.

Although the theory of the above mentioned three developments is well documented, their implementation in geotechnical engineering has been rather scarce, mainly because some of its concepts are not well understood by practitioners. In this way, there exists the need of a work that eluciates the theory behind these methodologies, and that serves as a framework for future applications of them in geotechnical engineering.

## 1.3 Aim and objectives

The main aim of this document is to elucidate a framework for conducting reliability assessments of geotechnical models, where epistemic and aleatory uncertainties can be propagated through the system and the probability of failure is computed in an accurate and efficient way, employing techniques of the current state-of-the-art.

To achieve this aim, the following objectives must be met:

- Learn the different representations of aleatory and epistemic uncertainty that can be used to model unknown geotechnical parameters, specially when the available information is scarce.
- Employ copula theory in order to model dependence among different geotechnical or engineering parameters.
- Learn state-of-the-art methods for the estimation of small probabilities of failure as an alternative to the costly Monte Carlo simulation method; in particular, a well-known method called subset simulation will be employed.
- Illustrate the applicability and advantages of the proposed methodology using several examples of the geotechnical engineering practice.

## 1.4 Document outline

The remaining of this document is organized as follows:

Chapter 2 presents some notions of probability theory and reliability theory that will be needed to understand the subsequent chapters. Chapter 3 is devoted to elucidate a new appealing simulation method for computing probabilities of failure, and exposes its role in geotechnical reliability analysis. Chapter 4 introduces copula theory and its role in dependence of random variables, with a practical emphasis on constructing multivariate probability models in geotechnics. Chapter 5 contains a succinct introduction to random set theory and to the different structures of information that it can handle in order to model aleatory and epistemic uncertainties. Chapter 6 groups the concepts of chapters 3, 4 and 5 to elucidate a comprehensive framework of reliability assessments of geotechnical models, which is illustrated through several examples. Finally, chapter 7 ends this document with the respective remarks and conclusions of this work.

# 2 Fundamentals of probability and reliability

The present chapter is devoted to expose some basic notions of probability theory and reliability analysis theory that will be needed hereinafter. In this way, this chapter is divided into two main sections. Section 2.1 deals with probability theory and is mainly based on [Ang and Tang \(2007\)](#) and [Kottegoda and Rosso \(2008\)](#). On the other hand, section 2.2 deals with reliability analysis theory and is mainly after [Melchers and Beck \(2018\)](#), [Ditlevsen and Madsen \(1996\)](#), [Lemaire \(2013\)](#) and [Ang and Tang \(1984\)](#).

## 2.1 Some concepts of probability theory

### 2.1.1 Axioms

Let us initially consider a probability space  $(\Omega, \mathcal{F}, P)$  where,  $\Omega$  is the sample space,  $\mathcal{F}$  is a  $\sigma$ -algebra of subsets  $\mathcal{A}$  of  $\Omega$ <sup>1</sup>, and  $P$  is a probability measure of the form  $P : \mathcal{F} \rightarrow [0, 1]$ , such that:

**Axiom 2.1.1.** *The probability of the event  $\mathcal{A}$  is a non-negative real number, for all  $\mathcal{A} \in \mathcal{F}$ :*

$$P(\mathcal{A}) \geq 0 \quad \forall \mathcal{A} \in \mathcal{F}$$

**Axiom 2.1.2.** *The probability of occurrence of at least one of all the possible events in the sample space is of 1.0:*

$$P(\Omega) = 1.0$$

**Axiom 2.1.3.** *If  $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \dots$  are mutually exclusive events, that is  $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$  for all  $i \neq j$ , then the probability of their joint occurrence is the sum of the individual probabilities of occurrence of each event:*

$$P\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \sum_{i=1}^{\infty} P(\mathcal{A}_i)$$

---

<sup>1</sup>A  $\sigma$ -algebra on the sample space  $\Omega$  is the collection of subsets  $\mathcal{A}$  of  $\Omega$ , which also includes  $\Omega$  itself and the empty subset. By definition, a  $\sigma$ -algebra is closed under complement, countable unions and countable intersections.

Axioms 2.1.1 to 2.1.3 are the basis of probability theory and are popularly known as the *Kolmogorov's axioms* after Andrey Nikolaevich Kolmogorov,<sup>2</sup> who stipulated them in 1933.

## 2.1.2 Random variables

Let  $(\Omega, \mathcal{F}, P)$  be a probability space. A *random variable (r.v.)*  $X$  is defined as a measurable function that maps the sample space  $\Omega$  to a measurable space, in most cases the real line  $\mathbb{R}$ :

$$X : \Omega \rightarrow \mathbb{R}$$

such that for each element  $\omega \in \Omega$  is associated an element  $x \in \mathbb{R}$ , that is  $\omega \mapsto x$ .

By mapping  $\Omega$  to  $\mathbb{R}$ , it is constituted a *measurable space* of the form  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , where  $\mathbb{R}$  is the real line and  $\mathcal{B}(\mathbb{R})$  is the Borel  $\sigma$ -algebra<sup>3</sup>. Furthermore, for each event  $F \in \mathcal{B}(\mathbb{R})$  a probability measure function  $P_X$  may be associated with the probability measure of the original space  $P$ , as:

$$\begin{aligned} P_X(F) &= P \circ X^{-1}(F) \\ &= P(X^{-1}(F)) \\ &= P\{\omega \in \Omega : X(\omega) \in F\} \end{aligned}$$

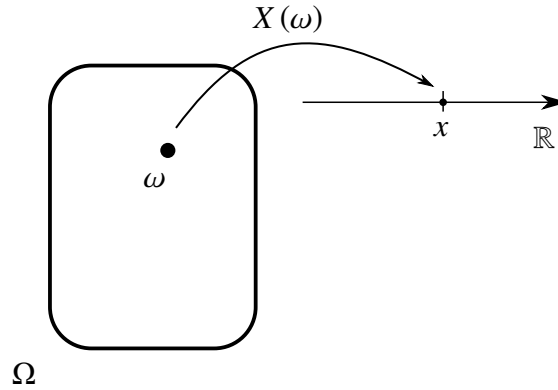
The triple  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$  conforms a new probability space which is the mathematical description of a random experiment  $\xi$ . Thus, a random variable is a function that maps an abstract probability space into a new probability space that mathematically describes the random experiment  $\xi$ , i.e.  $(\Omega, P) \xrightarrow{X} (\mathbb{R}, P_X)$ , where mathematical tools such as Riemann integration are well defined.

Figure 2-1 schematically depicts the transformation given by a random variable  $X$ .

Based on their range, random variables can be grouped in three major categories: discrete, continuous and mixed random variables. When the image of  $X$  is a countable finite set, then the random variable is said to be a *discrete random variable*; examples of them are the number of piles with excessive settlement after a load test or the number of failed slopes in a given area after an earthquake. On the other hand, when the image of  $X$  is an uncountable infinite set, then the random variable is said to be a *continuous random variable*; examples of them are the values of cohesion and friction angle of a given soil or the depth of the water table in a terrain. Finally, when  $X$  is neither discrete nor continuous, but rather a mixture of both, then the random variable is said to be a *mixed random variable*.

<sup>2</sup>Soviet mathematician (1903 - 1987) who devoted his life to the mathematics of probability theory, topology, intuitionistic logic, turbulence, classical mechanics, and computational complexity. In his book, *Foundations of the theory of probability*, published in 1933, lays the foundations of probability theory by the definition of the three basic statements of the discipline.

<sup>3</sup>The Borel algebra on  $\mathbb{R}$  is the smallest  $\sigma$ -algebra that contains all open sets (or all closed sets) of  $\mathbb{R}$



**Figure 2-1:** Random variable mapping from  $\Omega$  into  $\mathbb{R}$

### 2.1.3 Distribution functions

Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $X : \Omega \rightarrow \mathbb{R}$  a random variable, which represent a random experiment  $\xi$ . Bearing this in mind, the following definitions are presented:

**Distribution function:** A function of the form  $F(x) = P(X \leq x)$ , which meets the following properties:

1.  $F$  has limit at plus infinity, such that it satisfies:  $\lim_{x \rightarrow \infty} F(x) = 1$
2.  $F$  has limit at minus infinity, such that it satisfies:  $\lim_{x \rightarrow -\infty} F(x) = 0$
3.  $F$  is monotonically non-decreasing, that is,  $x \leq y$  implies that  $F(x) \leq F(y)$
4.  $F$  is right continuous, that is,  $\lim_{x \rightarrow a^+} F(x) = F(a^+)$ .

is known as the *cumulative distribution function* (CDF) of  $X$  or just the *distribution function* of  $X$ .

**Probability mass function:** A *probability mass function* (PMF) is defined as a function  $f(x) : \mathbb{R} \rightarrow \mathbb{R}$  that satisfies:

1.  $f(x) \geq 0$  for all  $x \in \mathbb{R}$
2.  $\sum_i f(x_i) = 1$ .

If  $X$  is a discrete random variable, then  $F(x)$  will be a piecewise constant function. Furthermore, PMFs and CDFs are related by the following relation:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

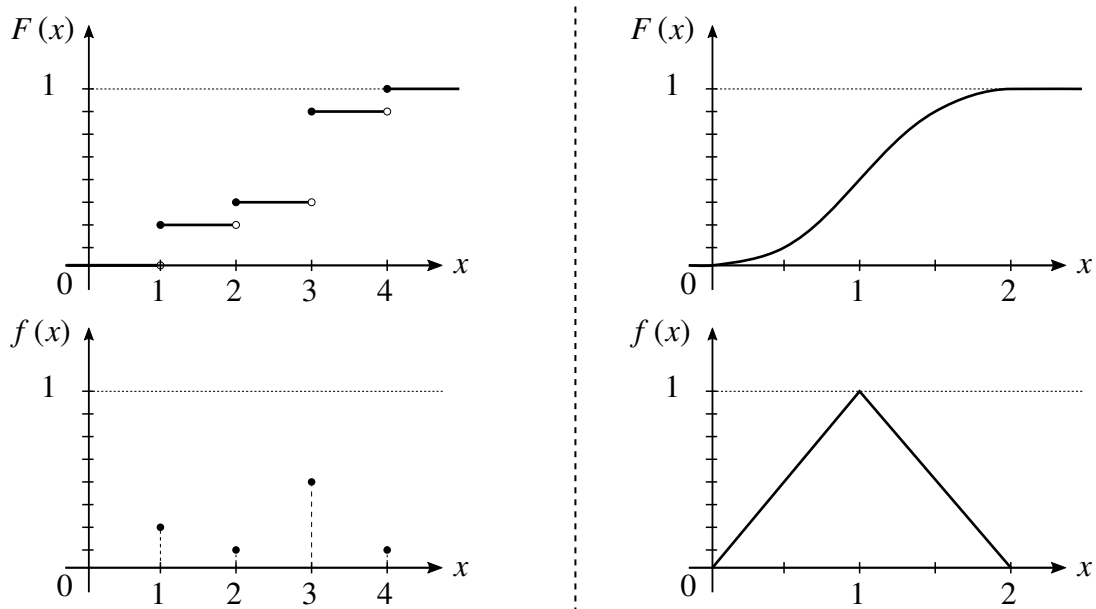
**Probability density function:** A *probability density function* (PDF) is defined as a function  $f(x) : \mathbb{R} \rightarrow \mathbb{R}$  that satisfies:

1.  $f(x) \geq 0$  for all  $x \in \mathbb{R}$
2.  $\int_{-\infty}^{\infty} f(x) dx = 1$ .

if  $X$  is a continuous random variable, then  $F(x)$  is a continuous function. Furthermore, PDFs and CDFs are related by the following relation:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x_i) dx.$$

To illustrate the above definitions, figure 2-2 depicts two random variables and their functions, either CDF or PMF/PDF.



**Figure 2-2:** Examples of discrete and continuous distributions. On the left side there is a discrete random variable represented by a CDF  $F(x)$  and a PMF  $f(x)$ . Right side of this figure contains the representation of a continuous random variable with a CDF  $F(x)$  and a PDF  $f(x)$ .

There exists a great variety of distribution functions in the literature. For a discrete case, it is worth mentioning some distributions such as the Bernoulli, Geometric and Poisson distributions. For a continuous case, it is worth mentioning some distributions such as the Normal, Log-normal, Gamma, Beta, Weibull, and Gumbel. In any case, an adequate introduction to these distributions may be found in standard references of probability theory, such as [Montgomery and Runger \(2010\)](#), [Ang and Tang \(1984\)](#), or [Kottegoda and Rosso \(2008\)](#).

### 2.1.4 Joint distribution functions

The concept of a random variable, as well as the concepts of distribution function and probability functions (either discrete, continuous or mixed) can be extended to several dimensions. Let us start by defining the function  $\mathbf{X} = (X_1, X_2, \dots, X_d)$  known as a random vector, such that  $\mathbf{X} : \Omega \rightarrow \mathbb{R}^d$  and  $d \geq 2$ .

In this way, each coordinate of a random vector  $\mathbf{X}$  corresponds to a random variable. Random vectors can be composed either of solely discrete, continuous, mixed random variables, or a mixture of all of them, although the latter case is rarely seen in geotechnical engineering practice. Additionally, it is worth mentioning that each random variable in  $\mathbf{X}$  may be independent of the others or has some degree of dependence with any, some, or all of the other random variables (this concept will be the main subject of section 4).

For the following definitions, let  $\mathbf{X} = (X, Y)$  be a bivariate random vector. Note that although a bivariate case is presented from simplicity, these definitions are easily extendable to more than two dimensions ( $d \geq 2$ ).

**Joint distribution function:** A function  $F(x, y) : \mathbb{R}^2 \rightarrow [0, 1]$ , where  $F(x, y) = P(X \leq x, Y \leq y)$ , that satisfies the following properties:

1.  $F$  has limit at plus infinity, such that:  $\lim_{x, y \rightarrow \infty} F(x, y) = 1$
2.  $F$  has limit at minus infinity, such that:  $\lim_{x \rightarrow -\infty} F(x, y) = \lim_{y \rightarrow -\infty} F(x, y) = 0$
3.  $F$  is monotonically non-decreasing in each variable, that is:

$$F(x_1, y) \leq F(x_2, y) \text{ when } x_1 \leq x_2$$

$$F(x, y_1) \leq F(x, y_2) \text{ when } y_1 \leq y_2$$

4.  $F$  is right continuous in each variable, such that:  $\lim_{x \rightarrow a^+} F(x, y) = F(a^+, y)$  and  $\lim_{y \rightarrow b^+} F(x, y) = F(x, b^+)$

is known as the *joint cumulative density function* of  $\mathbf{X}$ , or just the *joint distribution function* of  $\mathbf{X}$ .

**Joint probability density function:** A function  $f(x, y) : \mathbb{R}^2 \rightarrow [0, \infty)$ , non-negative  $f(x, y) \geq 0$  and integrable over  $\mathbb{R}^2$  as follows

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1,$$

is known as the *joint probability density function* of random vector  $\mathbf{X} = [X, Y]$  of continuous random variables. The following relations holds:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) \, du \, dv$$

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

**Marginal distribution function:** Functions  $F(x)$  and  $F(y)$ , defined as

$$F(x) = \lim_{y \rightarrow \infty} F(x, y) \quad \text{and} \quad F(y) = \lim_{x \rightarrow \infty} F(x, y)$$

are known as the *marginal distribution functions* of the random variables  $X$  and  $Y$ , respectively. In general  $F(x)$  and  $F(y)$  are different and turn out to be the individual distributions of the random variables  $X$  and  $Y$ . Note that, although it is possible to obtain  $F(x)$  and  $F(y)$  from  $F(x, y)$ , the opposite does not hold since from  $F(x)$  and  $F(y)$  there is no an unique definable  $F(x, y)$ .

**Marginal density function:** Functions  $f(x)$  and  $f(y)$ , defined as

$$f(x) = \int_{-\infty}^{\infty} f(x, y) \, dy \quad \text{and} \quad f(y) = \int_{-\infty}^{\infty} f(x, y) \, dx$$

are known as the *marginal density functions* of the random variables  $X$  and  $Y$ , respectively. Additionally, although from  $f(x, y)$  is possible to obtain  $f(x)$  and  $f(y)$ , the opposite does not hold, since from  $f(x)$  and  $f(y)$  there is no an unique definable  $f(x, y)$ ; this concept will be deepened in chapter 4.

**Conditional probability density function:** Let  $f(x, y)$  be the joint probability function of the random variables  $X$  and  $Y$ , and  $f(y)$  the marginal density function of  $Y$ , such that  $y \in \mathbb{R}$  and  $f(y) \neq 0$ . Then, the *conditional density function* of  $X$  given  $Y = y$ , can be defined as follows:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f(y)}$$

Note that in the above definitions  $\mathbf{X}$  is a vector of continuous random variables  $X$  and  $Y$ . If  $\mathbf{X}$  were a vector of discrete random variables  $X$  and  $Y$ , summations must be employed instead of integrals to define the *joint probability mass function*, *marginal probability mass function* or the *conditional probability mass function*.

### 2.1.5 Correlation among random variables

In the previous section, it was mentioned that the random variables that constitute a random vector  $X$  may be independent or have some degree of dependence to each other. A first approximation to this degree of dependence is given by the *correlation*. This property refers to degree to which a pair of random variables are related, and it is quite useful since it can indicate a predictive relationship between the variables of a model. In this way, correlation is a concept useful for describing simple relationships between basic variables of a model, without making a statement about cause and effect. Nonetheless, it is worth mentioning that dependence is a more general concept than correlation, since this latter represents just a first approximation to the dependence between two random variables.

Dependence is one of the most important and studied concepts for reliability assessments and probabilistic analyses (Jogdeo, 1982; Nelsen, 2007). Geotechnical engineering is not alien to this fact, and in practice, there exist different structures of dependence between many geotechnical parameters. For instance, there are relationships between shear strength parameters of the Mohr-Coulomb failure criterion ( $c - \phi$ ), between undrained shear strength ( $S_u$ ) and over-consolidation ratio (OCR), or even between ground motion parameters (GMPs). However, engineers may be more familiar with the concept of correlation than with the concept of dependence. In fact, geotechnical engineering is full of correlations among parameters, like those based on blow count of the SPT  $(N_1)_{60}$ .

The issue of correlation and dependence will be covered in more detail in chapter 4 of this document. In this chapter several measures of correlation are introduced.

### 2.1.6 Autocorrelation

In statistics, *autocorrelation* is the description of the relationship between observations of a random process over specific time periods. Informally, it is defined as the likeness between observations of the same process, as a function of the time difference or lag between them. Thus, autocorrelation constitutes a mathematical tool for finding repeated patterns or identifying the missing fundamental frequency in a signal (Wikipedia contributors, 2021a).

Let us consider  $X$  to be a random process, and  $i$  to be any point in time after the start of that process ( $i$  might be an integer or a real number considering whether a discrete-time or a continuous-time process, respectively). Then,  $X_i$  is the value (or realization) of the random process  $X$  at a time  $i$ . If the mean  $\mu_i$  and variance  $\sigma_i^2$  of the process at a time  $i$  are known, for each  $i$ , then the autocorrelation function between times  $i_1$  and  $i_2$  is:

$$R(i_1, i_2) = \frac{E[(X_{i_1} - \mu)(X_{i_2} - \mu)]}{\sigma^2}$$

where  $E[\cdot]$  is the expected value.

If the autocorrelation of a random process is measured as zero, it indicates that there is no correlation, and so, each of the observations is independent of the others. Conversely, if the autocorrelation of a random process tends to one, it indicates that the observations are serially correlated, and in consequence, future measurements are affected by past values. Essentially, a process that is serially correlated is not random, since it has a pattern.

### 2.1.7 Markov Chains

A *Markov Chain* is a mathematical system with a finite or countable number of states subject to transitions from one state to another, thus following a chain-like structure. It is defined as a stochastic process (Markov process) at a discrete-time, with a discrete space state, and fulfills the first order Markov property:

$$P(X_{i+1} = x_{i+1} | X_0 = x_0, X_1 = x_1, \dots, X_i = x_i) = P(X_{i+1} | X_i = x_i)$$

that is to say, the next state ( $x_{i+1}$ ) depends only on the current state ( $x_i$ ) and not on the other past events. In this sense, the prediction of the immediately following state depends only on the present event, and further information about past events does not lead to more accurate predictions.

Markov processes are named after its pioneer, Andrey Markov<sup>4</sup>, and they have been applied in many fields of knowledge, such as physics, biology, economy, engineering, statistics and among others ([Wikipedia contributors, 2021c](#)).

Transitions between states of a Markov chain are defined by the *transition probabilities*, denoted by  $P(s, t)$ ,  $P(s \rightarrow t)$  or simply  $P_{st}$ . Thus, a transition probability is defined as the probability of being at state  $t$  at time  $i + 1$ , given the current state  $s$  at time  $i$ .

When the transition probabilities are arranged in a stochastic matrix, then the *transition matrix*  $\mathbf{P}$  is consolidated. Each element of the transition matrix is nonnegative and the sum of each row is equal to one, that is  $\sum_t P_{st} = 1$  for all  $s$ . Thus, a Markov chain is completely characterized by the set of all its states and the respective transition probabilities.

Now, it is in our interest to know the probability of being in the state  $t$  at time  $i + 1$ . For this purpose, let

$$\boldsymbol{\pi}_s(i) = P(X_i = x_s)$$

denote the probability of state  $s$  at time  $i$ , and also let  $\boldsymbol{\pi}_s(i)$  denote the row vector of the state transition matrix at the step  $i$ .

Additionally, every chain has to initialize by specifying a starting vector  $\boldsymbol{\pi}(0)$ , where all its elements have a probability of occurrence equal to zero, except one element that has a probability of occurrence equal to 1, which corresponds to the initial state.

<sup>4</sup>Russian mathematician (1856 - 1922) widely known by his contributions on the theory of stochastic processes. His first results on Markov processes and Markov chains were published in 1906.

In this way, the total probability theorem can be applied to obtain the probability of the chain being at state value  $x_t$  at time  $i + 1$ , as follows:

$$\begin{aligned}\pi_t(i + 1) &= P(X_{i+1} = x_t) \\ &= \sum_r P(X_i = x_r) P(X_{i+1} = x_t | X_i = x_r) \\ &= \sum_r \pi_r(i) P_{rt}\end{aligned}$$

By employing the transition matrix  $\mathbf{P}$ , a more compact expression can be obtained, as follows:

$$\boldsymbol{\pi}(i + 1) = \boldsymbol{\pi}(i) \mathbf{P}$$

Analogously, using the last state:

$$\boldsymbol{\pi}(i) = \boldsymbol{\pi}(i - 1) \mathbf{P} = (\boldsymbol{\pi}(i - 2) \mathbf{P}) \mathbf{P} = \boldsymbol{\pi}(i - 2) \mathbf{P}^2$$

And by continuing with this sequence:

$$\boldsymbol{\pi}(i) = \boldsymbol{\pi}(0) \mathbf{P}^i \tag{2-1}$$

Finally, when the probabilities of being in any particular state are independent of the initial condition, the Markov chain is said to have achieved a *stationary distribution*  $\boldsymbol{\pi}^*$ . In other words, a stationary distribution of a Markov chain is some vector  $\boldsymbol{\pi}^*$  such that  $\boldsymbol{\pi}^* = \boldsymbol{\pi}^* \mathbf{P}$ . Thus, over a long run, no matter the starting state  $\boldsymbol{\pi}(0)$ , the probability of being at state  $s$  is  $\pi_s^*$  for all  $s$  of the Markov Chain.

At this point, it is worth mentioning some properties that will be needed later and which also help to clarify the above concepts. Let us consider the following definitions (Uribe, 2011):

**Definition 2.1.1.** A Markov chain is said to be *stationary* when the time has no implications on the random nature of the stochastic process (Markov process), so all states  $x_s$  are independent of time. Thus, a Markov chain has reached the stationary distribution when the probabilities of all its states are independent of the actual starting value.

**Definition 2.1.2.** A Markov chain is said to be *ergodic* when the number of iterations on the chain tends to infinity ( $i \rightarrow \infty$ ). Note that a Markov chain can only have one stationary distribution, and additionally, all ergodic processes are stationary.

**Definition 2.1.3.** A Markov chain is said to be *homogeneous* when the transition probabilities do not change in the progression of state transitions. In other words, a Markov chain is homogeneous if  $\mathbf{P}$  is constant in time.

**Definition 2.1.4.** A Markov chain is said to be *irreducible* when for two states  $s$  and  $t$  of it, there exists a probability of not entering to state  $t$  given the current state  $s$ . Thus, all states of the chain are connected with each other.

**Definition 2.1.5.** A Markov chain is said to be *aperiodic* when the chain does not fall into some kind of periodic cycle, i.e. the number of steps to move from a state  $s$  to a state  $t$  is not necessary to be multiple of some integer.

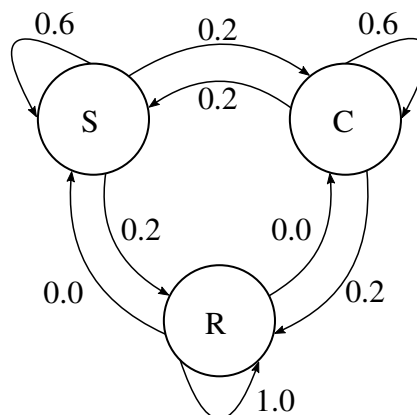
**Definition 2.1.6.** A Markov chain is said to be *reversible* when it satisfies the so called reversibility condition:

$$P_{tr}\pi_t^* = P_{rt}\pi_r^*$$

In this sense, if a certain time  $i$  the direction of the Markov chain is reversed, the process behavior will be preserved, i.e. the resulting chain will be statistically indistinguishable from the original one.

To illustrate these, the example presented by Uribe (2011) will be replicated below. On the other hand, readers interested in a more comprehensive introduction to the theory of Markov chains, and in further examples, can refer to Walsh (2004), Ash (2008), Benjamin and Cornell (1981) or Wikipedia contributors (2021b), and the references therein.

**Example 2.1.1.** (adapted from Uribe (2011)) Let us consider three weather conditions, namely sunny (S), cloudy (C) and rainy (R) and the probabilities of transition among them, as stated in figure 2-3.



**Figure 2-3:** Possible weather states and probabilities of transition among them

Based on figure 2-3 the following transition matrix can be constituted. In this matrix state 1 is sunny, state 2 is cloudy, and state 3 is rainy:

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Bearing this information in mind, the following questions are required to be solved:

- Given a sunny initial state (state 1) at time zero, what is the probability of the rainy state (state 3) at time 2
- For a long time, what is the probability of the condition being in each one of the three states

By employing the equation (2-1), the first question can be solved as follows:

$$\boldsymbol{\pi}(i) = \boldsymbol{\pi}(0) \mathbf{P}^i$$

$$\boldsymbol{\pi}(2) = \boldsymbol{\pi}(0) \mathbf{P}^2$$

but since  $\boldsymbol{\pi}(0) = [1 \ 0 \ 0]$  because sunny (S) is the initial state, equation (2-1) can be solved as follows

$$\boldsymbol{\pi}(2) = [1 \ 0 \ 0] \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}^2$$

$$\boldsymbol{\pi}(2) = [0.4 \ 0.24 \ 0.36]$$

Then, the probability of being at state 3 (rainy) at time 2 is equal to 36%.

With respect to the second question, let us find the probabilities of each state at times 5, 15 and 30, as follows

For  $i = 5$ :

$$\boldsymbol{\pi}(5) = \boldsymbol{\pi}(0) \mathbf{P}^5$$

$$\boldsymbol{\pi}(5) = [0.17 \ 0.16 \ 0.67]$$

for  $i = 15$ :

$$\boldsymbol{\pi}(15) = \boldsymbol{\pi}(0) \mathbf{P}^{15}$$

$$\boldsymbol{\pi}(15) = [0.0176 \ 0.0176 \ 0.9648]$$

and for  $i = 30$ :

$$\boldsymbol{\pi}(30) = \boldsymbol{\pi}(0) \mathbf{P}^{30}$$

$$\boldsymbol{\pi}(30) = [0.00062 \ 0.00062 \ 0.99876]$$

It is evident that for a long run, the probabilities of states 1 and 2 (sunny and cloudy, respectively) will be zero, and the probability of state 3 (rainy) will be one. Note that for a long time the chain will reach its stationary distribution, which is obviously independent of the initial state  $\pi(0)$ .

## 2.2 Introduction to reliability analysis

### 2.2.1 The safety challenge

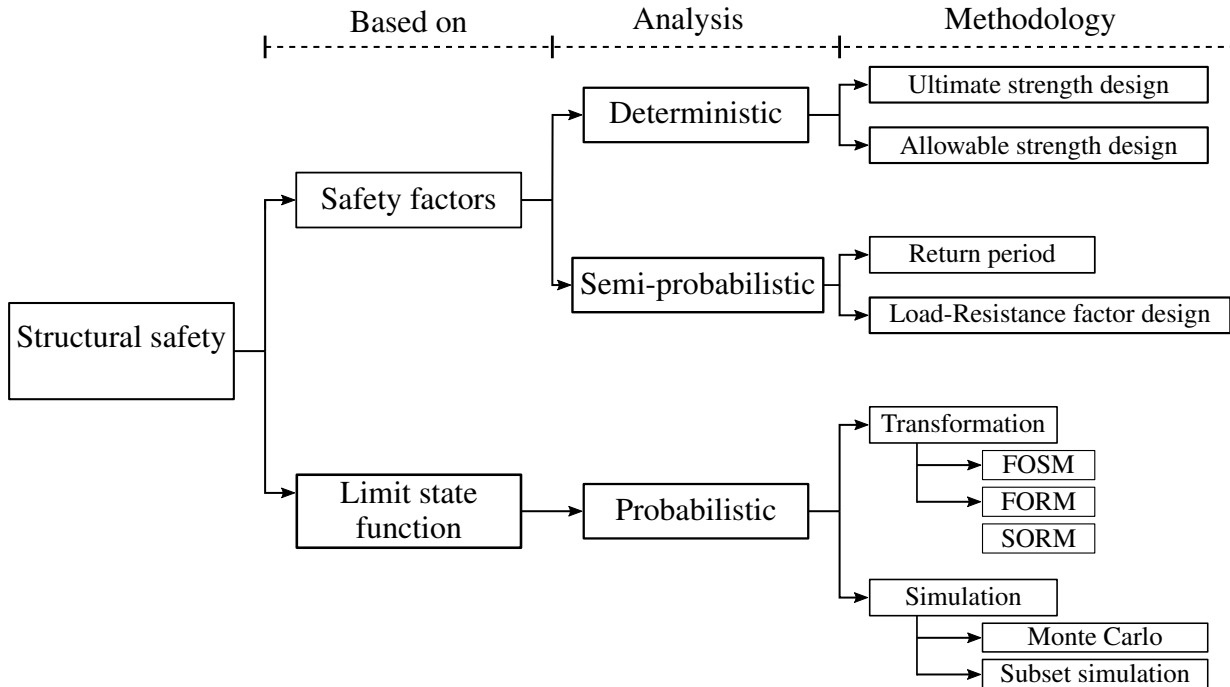
Bearing in mind the inherent uncertainties in all geotechnical designs-interventions, as stated in section 1.1, the question now is how to ensure their safety. In other words, despite both aleatory and epistemic uncertainties, all geotechnical designs must guarantee a suitable performance during its estimated service life, preserving goods, services, and above all, human life.

Recall that in geotechnical engineering aleatory uncertainty is considerable and non-reducible and although epistemic uncertainty may be reduced with additional information, the truth is that data is limited due to reduced budgets or haste in the project schedule. Under these circumstances, all designs must be evaluated in light of a safety margin or reliability as a measure to ensure their faultless performance.

There exist several perspectives with which the challenge of structural safety can be addressed. Most of these approaches are based on limit states, i.e. the state of impending failure, beyond which the structure serviceability ceases. Figure 2-5 shows a conceptual map that, in a general way, summarizes the type of analyses and methodologies that can be employed to evaluate the reliability of a design.

It is worth mentioning that a limit state does not necessarily imply failure. Limit states can be divided into *ultimate* and *serviceability limit states*. On the one hand, ultimate limit states do imply failure, collapse, or a loss of integrity of the structure, and hence they are not reversible; an example of its violation is the collapse of the Nicoll Highway in Singapore (see [Puzrin et al., 2010](#); [Whittle and Davies, 2006](#)). On the other hand, serviceability limit states refer to the threshold between an acceptable and an undesirable state of the structure under normal use, and most of the time they are reversible or remediable; an example of its violation are excessive settlements, such as the ones reported in the Palacio de Bellas Artes in Mexico City (see, e.g., [Lambe and Whitman, 1991](#))

In the remaining of this section, the different safety analysis exposed in figure 2-5 will be briefly introduced.



**Figure 2-4:** Structural safety approaches (adapted from [Hurtado \(2004\)](#) and [Marek et al. \(1996\)](#)).

### 2.2.1.1 Deterministic analysis

Traditionally, deterministic analysis has been the most employed analysis for evaluating safety of structures, given its simplicity and familiarity for many engineers. Deterministic analysis is based on the concept of *safety factor*, which is defined as the following ratio ([Ditlevsen and Madsen, 1996](#)):

$$FS = \frac{R}{S} \quad (2-2)$$

where  $R$  is the calculated carrying capacity (resistance), and  $S$  the corresponding load effect (stress). Note that  $S$  is made up of the effects of one or more applied loads, such as dead loads, water loads, earthquakes loads, etc.

Theoretically, a geotechnical structure can be classified into one of three different states from a deterministic analysis. If  $FS > 1.0$  then  $R > S$ , and therefore the design is safe. If  $FS < 1.0$  then  $R < S$ , and therefore there is failure. And finally, if  $FS = 1.0$  then  $R = S$ , and therefore the structure is in a state of imminent failure.

At first glance, one can expect that the greater  $FS$  in equation (2-2), the greater the reliability of the model. Naturally, for a given definition of  $R$ , and hence of  $S$ , an increase of  $FS$  reflects an increase of safety. Nonetheless, the safety factor highly depends on the definition of the resistance  $R$ , which may contain a certain degree of arbitrariness. This arbitrary nature of the definition of  $R$  makes minimum values of  $FS$  difficult to define in the context of codes and norms. Necessarily,

a specification of the value of a safety factor  $FS$  in a code or norm must be accompanied by specifications of its derivations, or the formulas for defining its respective resistance. Even so, in some problems such as the ones of soil mechanics, the resistance is not definable in a clear way as a single scalar quantity (see [Ditlevsen and Madsen, 1996](#), chap. 2). This problem is known as *lack of invariance* of the safety measure. Ideally, a safety measure should not be dependent on the way the resistance  $R$  and the load  $S$  are defined ([Melchers and Beck, 2018](#)).

Codes and norms contain specifications of the minimum value of  $FS$  for common and simple models in geotechnics, which do not have a high degree of uncertainty. The magnitude of the minimum safety factors stems from past experiences in similar models and structures, but in turn, it is conditioned by social factors, economy, and in general, particular conditions of each region or country. Extremely high values of  $FS$  will lead to too conservative models that will be too expensive and whose execution time is quite long. Contrarily, low values of  $FS$  that tend to 1.0 will lead to unconservative models and structures that will probably fail if the uncertainty is considerable. Consequently, the selection of the minimum values of  $FS$  is the responsibility of a code committee that must take into account all these considerations ([Ditlevsen and Madsen, 1996](#)).

Note that safety factors ignore much information that may be available about uncertainties in  $R$  and  $S$ . In essence, the deterministic analysis does not consider uncertainty explicitly, and the final value of a factor of safety will be the result of the assumptions and considerations made for its computation. Some other approaches can be employed to integrate uncertainties with deterministic analysis, such as the *partial safety method* (see, e.g., [Ditlevsen and Madsen, 1996](#); [Melchers and Beck, 2018](#)), or an approach in which fractiles of  $R$  and  $S$  are taken for computing the factor of safety. Nevertheless, in any case, explicit handling of uncertainty is only achieved with other approaches of structural safety, especially those focused on reliability estimates such as the probabilistic analysis.

### 2.2.1.2 Semi-probabilistic analysis

Semi-probabilistic analysis can be understood as a combination of a deterministic analysis and a probabilistic analysis. In a semi-probabilistic approach, randomness is included in the definition of a design load and-or a reduced capacity, and the rest of the study is carried out as deterministic, so the safety factor is computed as usual. In this sense, semi-probabilistic analyses constitute an improvement over deterministic analyses, but at the same time, randomness is not considered in its entirety, and it is not propagated in the model, so this approach is only partially probabilistic.

One example of semi-probabilistic approaches are those based on the methodology of the *return period*. In this methodology, loads due to natural phenomena, e.g., winds, waves, storms, floods, earthquakes, etc. are taken as a random process in time, characterized by the return period. This return period refers to the expected time between two successive independent events. In this sense, a design load can be stated as the load corresponding to a given return period, e.g., the peak ground acceleration for a 475 years earthquake. Nonetheless, the methodology of the return period

depends on the time scale, the possible occurrence of more than one event within a time period is ignored, and the probability distribution of the load given by the natural phenomena is not considered. Thus, the return period is a probabilistic measure only in terms of time, but the actual magnitude of the loadings and their interaction with resistance is ignored (Melchers and Beck, 2018).

Other examples of semi-probabilistic approaches are the ones based on the methodology of *Load Resistance Factor Design* (LRFD). In this methodology, the magnitude of each load and the value of the resistance are affected by factors of design, which in turn depend on the variance of the loads and resistance, respectively. Values of the factors of design are based on statistical records and past experiences, and usually they are already formulated in the different codes and guides. After applying these factors to the loads and resistance, the rest of the analysis is carried out deterministically.

### 2.2.1.3 Probabilistic analysis

Probabilistic analysis is a reliability approach intended to estimate the probability of either an ultimate or a serviceability limit state violation for a structure or model, at any stage during its life. Probabilistic assessments of reliability are an extension of deterministic analysis, where any uncertainty about a deterministic quantity is taken into account explicitly. While in deterministic analysis the strength and applied loads are assumed to be known and there is no place for uncertainty, in probabilistic analysis both resistance and applied loads are taken as random variables that can be represented by a distribution function (Melchers and Beck, 2018).

At this point, it is important to introduce the *limit state function*  $G$ , also known as the *performance function*. This function represents the limit state of the models and divides the entire domain of the vector of random variables  $\mathbf{X} \in \mathbb{R}^d$  into two subdomains, the *failure domain*  $F_D$  and the *safe domain*  $S_D$ , defined as follows:

$$\begin{aligned} F_D &= \{\mathbf{x} \in \mathbb{R}^d : G(\mathbf{x}) \leq 0\} \\ S_D &= \{\mathbf{x} \in \mathbb{R}^d : G(\mathbf{x}) > 0\} \end{aligned}$$

In this way, the main aim of the probabilistic assessment of reliability is to compute the probability of limit state violation, or the probability of  $F_D$ , which is better known as the *probability of failure*  $P_f$ . This probability will be the probability of an undesirable behavior and it considers the aleatory uncertainty of the basic variables of the studied model.

When the model can be reduced to a resistance and load effects, the probability of failure  $P_f$  is computed by the following integral (Melchers and Beck, 2018):

$$P_f = P(G(R, S) \leq 0) = \iint_{F_D} f_{RS}(r, s) dr ds \quad (2-3)$$

where  $R$  is the random variable that represents resistance,  $S$  the random variable that represents load effects,  $F_D$  is the failure domain, and  $f_{RS}(r, s)$  the bivariate probability density function of  $R$  and  $S$ .

Note from equation (2-3) that  $G(R, S)$  requires an expression for its evaluation. In most cases, the performance function is presented as the difference of the resistance and load effects, that is to say

$$G(R, S) = R - S;$$

nonetheless, this is not the unique expression and other equations can be employed, for example (Melchers and Beck, 2018):

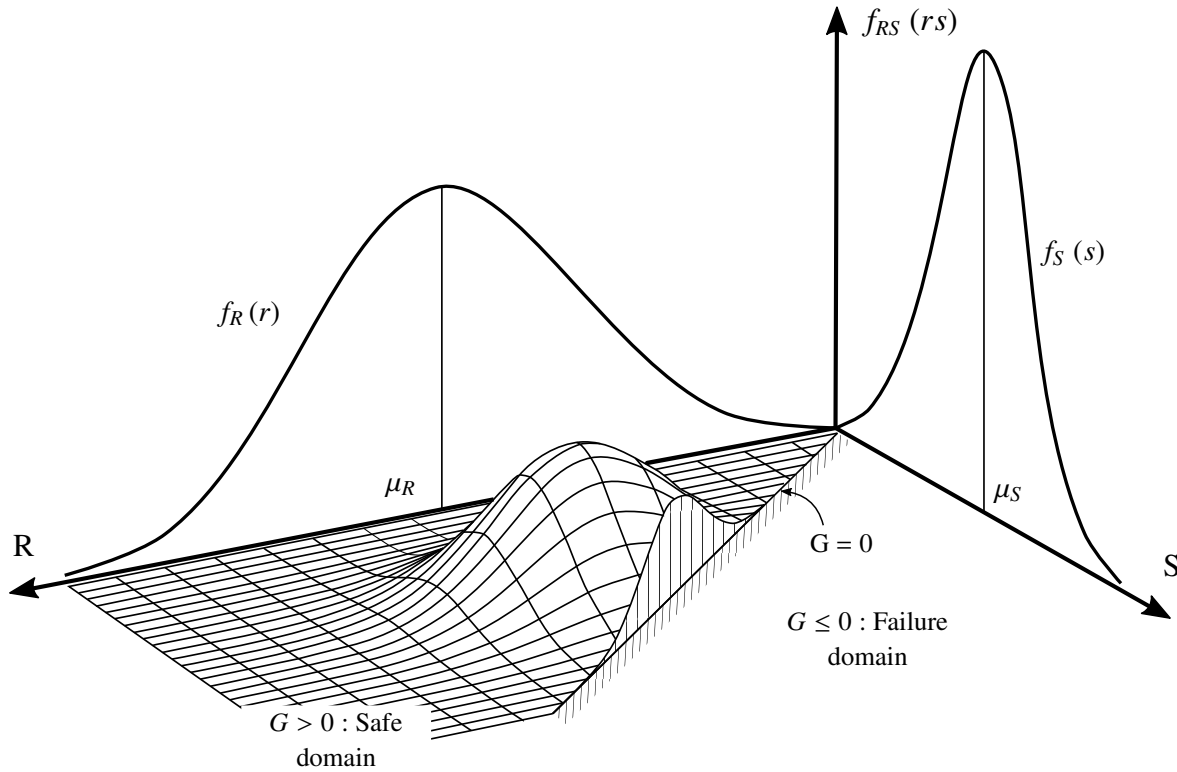
$$G(R, S) = \frac{R}{S} - 1.0,$$

which is the way of express the factor of safety  $FS = R/S$  in a probabilistic way.

For a better understanding, equation (2-3) is schematically represented in figure 2-5. Note in figure 2-5 that the marginal density functions of  $R$  and  $S$  are also plotted,  $f_R(r)$  and  $f_S(s)$ , respectively. Additionally,  $\mu_R$  and  $\mu_S$  represent the expected values of  $R$  and  $S$ , respectively. These two values will lead to the basic safety factor by replacing them in equation (2-2). However, it is evident from figure 2-5 that by employing a fully deterministic analysis much uncertainty is being ignored, and the final factor of safety may not reflect the true state of safety of the analyzed structure.

Nevertheless, most of the time, equation (2-3) is mathematically untractable. In the first place, it is not always possible to reduce a system to be represented by its resistance  $R$  and load effects  $S$ , being this problem particularly evident in soil and rock mechanics. In this regard, note that resistance  $R$  depends on material properties and dimensions, while load effects  $S$  depend on the applied loads, material densities, and perhaps structural dimensions. On the other hand, to solve equation (2-3) some assumptions are made, such as the independence of  $R$  and  $S$  or its Gaussianity, but these assumptions are almost always unsubstantiated. For example,  $R$  and  $S$  are not independent when some loads act opposite to failure, or when some parameters affect both  $R$  and  $S$ . This is the case of the overturning analysis of a retaining wall, where the soil unit weight affects both  $R$  and  $S$ , or in the case of an isolated footing with an overload adjacent to it. Under these circumstances, a more general formulation is required to compute the probability of failure of a system. In this regard, performance function  $G$  has to be expressed in terms of the basic variables of the analyzed system.

Basic variables refer to the fundamental variables that control the behavior of the studied system, and additionally, these are taken as random variables when there is randomness in their characterization. Examples of basic variables that can be considered random are soil shear strength parameters, undrained shear strength, strike and dip, etc. Therefore, when the geotechnical models are expressed through their basic variables, then the *generalized limit state function*  $G(\mathbf{X})$  must be employed, where  $\mathbf{X}$  is the vector of basic variables that compound the problem. Unlike  $G(R, S)$  that



**Figure 2-5:** Scheme of the bivariate probability function of a model reduced to resistance forces  $R$  and load effects  $S$ , with the details of the univariate PDF of  $R$  and  $S$ , the limit state function  $G$ , and the safe and failure domains.

can be expressed as  $R - S$  (or any of its alternatives),  $G(\mathbf{X})$  is expressed in terms of the equations that represent each problem and particular analysis.

When the generalized limit state function  $G(\mathbf{X})$  is employed, the probability of failure  $P_f$  is defined as (Alvarez et al., 2018):

$$P_f = \int \cdots \int_{\mathbb{R}^d} \mathbb{I}[G(\mathbf{X}) \leq 0] dF_{\mathbf{X}}(\mathbf{x}) \quad (2-4)$$

where  $F_{\mathbf{X}}(\mathbf{x})$  is the CDF of the input variables, and  $\mathbb{I}[\cdot]$  stands for the *indicator function*, which is a function that takes a value of 1 if its respective condition is met, or 0 otherwise. Note that  $\mathbf{X}$  represents the  $d$ -dimensional vector of random variables and  $\mathbf{x}$  is a  $d$ -dimensional point in the basic variable space.

If  $F_{\mathbf{X}}(\mathbf{x})$  is sufficiently differentiable, then the joint PDF  $f_{\mathbf{X}}(\mathbf{x})$  exists, and hence equation (2-4) can be written as (Melchers and Beck, 2018):

$$P_f = P(G(\mathbf{X}) \leq 0) = \int \cdots \int_{G(\mathbf{X}) \leq 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \quad (2-5)$$

which can also be expressed as:

$$P_f = \int \cdots \int_{\mathbb{R}^d} \mathbb{I}[G(\mathbf{X}) \leq 0] f_X(\mathbf{x}) d\mathbf{x} \quad (2-6)$$

Although equation (2-5) is more general and explicit than equation (2-3), its solution is rarely obtained analytically. Furthermore, complexity of equation (2-5) increases when the number of random variables  $d$  increases, when there is some dependence among the random variables of  $\mathbf{X}$ , or when the performance function  $G(\mathbf{X})$  is strongly non-linear. It is also worth mentioning that equation (2-5) only takes into account aleatory uncertainty (randomness) but not epistemic uncertainty, which may be transcendental in some problems with limited knowledge/data, such as the ones of geotechnical engineering.

Next section introduces some methodologies to approximate  $P_f$ , but the main approach to this end is explained in chapter 3, which is dedicated to simulation methods and in particular to a novel methodology called the subset simulation algorithm

## 2.2.2 Computing probability of failure

The main aim of reliability analysis is to characterize all the uncertainties of a system and compute its probability of failure through equation (2-5). However, equation (2-5) seldom has an analytical solution due to its complexity, which is increased with models in high dimensions, when there is dependence between the basic variables, and/or when there is non-linearity on the performance function. Under these circumstances, analysts must appeal to methods that accurately approximate equation (2-5).

This section introduces the basic concepts of some popular methods that approximate equation (2-5). These methods are classified into three major categories: numerical integration methods, transformation methods, and simulation methods. Our principal interest is in simulation methods, although in this section only Monte Carlo simulation is introduced. A modern and appealing simulation method called subset simulation will be the main subject of chapter 3.

### 2.2.2.1 Numerical integration

Analytical solution of equation (2-5) is possible only for some particular cases of limited practical interests. However, there exist some numerical integration methods in the literature for the approximation of problems of this nature that work in low dimensions such as the trapezoidal rule, Simpson's rule, Laguerre-Gauss or Gauss-Hermite quadrature formulae (see, e.g., [Chapra et al., 2012](#)). Nonetheless, with large dimensions (say  $d \geq 5$ ) numerical integration methods are seldom useful due to the growth of round-off errors and excessively long computation times.

Sometimes, it is possible to reduce the complexity of equation (2-5) and apply a numerical integration method. For example, if the performance function  $G(\mathbf{X})$  is a linear function, it is possible to

express equation (2-5) as a series of single integrals and apply some numerical integration to approximate them. Nonetheless, the approximation still requires extensive numerical work or further simplifications to be practical.

Thus, although at first glance numerical integration methods would be suitable for computing equation (2-5), the reality is that their applicability is rather limited. Instead, some other approaches based on transformations or simulations should be employed.

### 2.2.2.2 Transformation methods

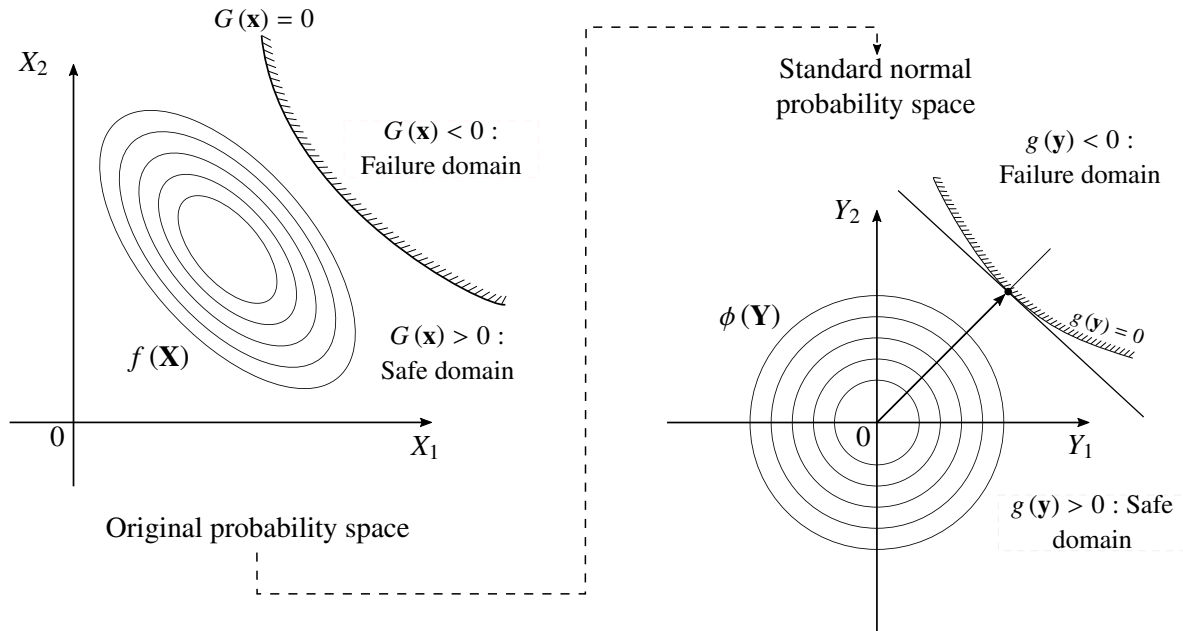
Transformation methods seek to sidestep the multiple integrals of equation (2-5) by applying a transformation that directly allows assessing reliability. Specifically, these methods transform the original probabilistic space into a standardized normal probability space. Working in this new space has many advantages related to the remarkable properties of the standardized multivariate normal distribution.

In general, all methods that employ this class of transformation follows a similar procedure to assess reliability. Basically, these methods transmute the original random variable  $X$  into the standard normal random variable  $Y$ . Evidently, if the probability space is transformed, so must be the performance function; thus, the original limit state function  $G(X) = 0$  must be transformed into a new function given by  $g(Y) = 0$ . Figure 2-6 illustrates this procedure. It is worth noting that in Figure 2-6,  $\Phi$  represents the standard normal CDF, and  $\phi$  represents the standard normal PDF, so it must not be confused with soil friction angle, which usually represented by the same symbol.

Transformation methods have a major drawback: the joint probability function of the random variables involved in the model must be known with certainty. This level of accuracy is rarely achieved in geotechnical engineering practice due to the lack of information, and hence, approximations to the joint probability functions must be made. Many of these approximations employ the marginal probability function of each random variable and the correlation matrix that relate them, which is more feasible in practice. A good introduction to these methods, and their assumptions, can be found in [Der Kiureghian and Liu \(1986\)](#).

In this sense, although transformation methods have been quite popular for assessing reliability, their use implies several approximations that the analyst should be aware of. When these methods are employed without knowing their hypothesis, estimates of reliability may be misleading. In several cases, these methods constitute a rough estimative of equation (2-5), so it is evident the need for a more general and accurate procedure.

Further details about the theory of transformation methods, and their algorithms, can be found in other specialized references, such as [Melchers and Beck \(2018\)](#). An introduction of the employability of transformation methods in geotechnical engineering can be found in [Baecher and Christian \(2005\)](#).



**Figure 2-6:** Graphical representation of reliability assessments based on the approach of transformation to the standard normal space.

### 2.2.2.3 Simulation methods

Simulation can be defined as a numerical technique for conducting experiments on a computer, where the behavior of real systems is represented through certain types of mathematical and logical models (Naylor et al., 1966). In most cases, simulation methods may be appropriate to study complex problems without an analytical solution, such as in equation (2-5). Thus, in our context, simulation methods can be employed to assess the reliability of geotechnical systems by computing their probability of failure. Particularly, Monte Carlo simulation (MCS) is widely accepted, since it can solve multidimensional integrals with complicated boundary conditions.

Monte Carlo Simulation was originally applied by John Von Neumann, Stanislaw Ulam, and Nicholas Metropolis as a practical solution to complex problems while they were working on nuclear weapon projects during the second world war. Monte Carlo method was initially a code name that comes from an analogy to the gambling casinos in the city of Monte Carlo at Monaco, but after the war finished, this methodology became popular, and so its initial name. MCS is mainly useful for solving three types of problems: optimization, numerical integration, and sampling from a probability distribution. It is considered a fairly general and robust methodology, valid in many areas of knowledge, such as chemistry, physics, mathematics, engineering, etc.

In the context of reliability analysis, MCS serves to propagate uncertainties through a mechanical system, in order to obtain an estimative of its behavior. In this sense, MCS serves to approximate

the  $P_f$  of a model, and hence its reliability. For this purpose, MCS employs an unbiased estimator<sup>5</sup> of equation (2-6), which is expressed through the following equation (Melchers and Beck, 2018, chap. 3):

$$\hat{P}_f \approx \frac{1}{N} \sum_{j=1}^N \mathbb{I}[G(\mathbf{x}_j) \leq 0] \quad (2-7)$$

which is known as the crude Monte Carlo approximation.

In equation (2-7),  $\mathbf{x}_j$  is a realization of the random vector  $\mathbf{X}$ , and  $N$  is the number of runs of the simulation. Thus, the  $\hat{P}_f$  estimate through equation (2-7) can be interpreted as a statistical averaging of the indicator function, such that for every  $\mathbf{x}_j$  with  $j = 1, 2, \dots, N$ , the following condition is evaluated.

$$\begin{aligned} \text{Failure region:} & \quad \mathbb{I}[\mathbf{x}_j \in F] = 1 & \quad \text{since } G(\mathbf{x}_j) \leq 0 \\ \text{Safe region:} & \quad \mathbb{I}[\mathbf{x}_j \in F] = 0 & \quad \text{since } G(\mathbf{x}_j) > 0 \end{aligned}$$

From equation (2-7), it is worth highlighting two important remarks, one related to the sampling and another related to the accuracy of the approximation of  $P_f$ .

On the one hand,  $\mathbf{x}_j$  is termed as a pseudo-random sampling since it is not totally random. There is an underlying formula or algorithm which generates  $\mathbf{x}_j$ . However, for most practical situations  $\mathbf{x}_j$  is indistinguishable from a true random sample from  $\mathbf{X}$ . In the literature, there exist several algorithms for sampling random numbers from a probability distribution, such as the inverse transform, acceptance-rejection, Box-Muller, Latin hypercube sampling, ziggurat, etc. Exposure of these methods is out of the scope of this document, so readers are referred to Ross (2012) or Rubinstein and Kroese (2016) for a detailed description. Furthermore, most programming languages have routines for sampling pseudo-random numbers from the most common probability distributions. Additionally, note that there may be some dependence among the random variables of  $\mathbf{X}$ , that must be considered when sampling.

On the other hand, the number of simulations  $N$  is directly related to the accuracy of the final estimate of  $P_f$ . As  $N$  approaches to infinity,  $\hat{P}_f$  converges to its true value,  $P_f$ . But on the contrary, when  $N$  is not large enough then the estimate of  $P_f$  may be a vague approximation to the real value. Even nowadays, the minimum sample size  $N_{min}$  of a MCS is a widely debated issue. In the literature, there are several equations aimed to obtain an initial estimate of the minimum sample size of a MCS, such as:

$$N_{min} \approx \frac{100}{P_f} \quad (2-8)$$

<sup>5</sup>The bias of an estimator refers to the difference between the expected value of the estimator and the true value of the parameter that is being estimated. Thus, an unbiased estimator is an estimator with bias equal to zero

or many other found in specialized references (see, e.g., [Ang and Tang, 1984](#); [Marek et al., 1996](#)). However, although rules of this nature are often useful, they say little or nothing about the achieved accuracy of any particular MCS.

In any case, it is a good practice to plot progressive results of the estimate  $P_f$  obtained through equation (2-7), and their estimated coefficient of variation (c.o.v.), defined as ([Zuev et al., 2012](#)):

$$\delta(\hat{P}_f) = \sqrt{\frac{1 - P_f}{NP_f}}$$

Thus, it will be noted that as the number of simulations increases, the values of  $\hat{P}_f$  (and hence their  $\delta(\hat{P}_f)$ ) reach a stationary value. At this point, it could be said that the number of samples is sufficient; otherwise, the number of samples should increase.

Nonetheless, a significantly large number of simulations leads to one of the major disadvantages of MCS. Although MCS is a robust methodology to face problems in high-dimensions and with complex limit state functions, it is excessively costly in terms of computer resources. In other words, MCS is not an efficient methodology to compute equation (2-5). Running a complex model many times is quite demanding, and it may exceed even the current computational capacities.

The drawback of efficiency becomes more evident when the failure probabilities to be computed are too small (say  $P_f \leq 10^{-3}$ ) since these require a large number of simulations to achieve an accurate estimate. This is the case of probabilities of failure associated with rare events, such as earthquakes or extraordinary rains. By employing a crude MCS, a sample from the failure domain is only obtained after obtaining a large number of samples in the safe set.

However, a process of this nature is quite inefficient since our real interest is the failure domain. Thus, for complex models associated with a small probability of failure, a crude MCS may be unfeasible. Due to this, other simulation methods, mostly based on the concept of MCS, have been developed seeking to overcome the drawback of efficiency of the crude MCS. Among all these methods, it is worth mentioning the *importance sampling* and the *directional simulation*. An explanation of these alternatives is out of the scope of this document, but interested readers are referred to [Melchers and Beck \(2018\)](#). Nonetheless, we are interested in a relatively new simulation methodology that has all the advantages of the MCS, but at the same time overcomes the drawback of efficiency. This methodology will be the main issue of chapter 3.

## 3 Subset simulation for assessing geotechnical reliability

As stated in section 2.2.2, the main aim of a reliability analysis is to accurately estimate  $P_f$  from equation 2-5. Nonetheless, this computation is not trivial since, in most practical cases, equation 2-5 has no analytical solution. In this regard, a direct numerical integration of equation 2-5 is seldom feasible, and a transformation of its multiple integrals to a standardized normal space carries out several hypotheses and approximations that, in most cases, have no foundations.

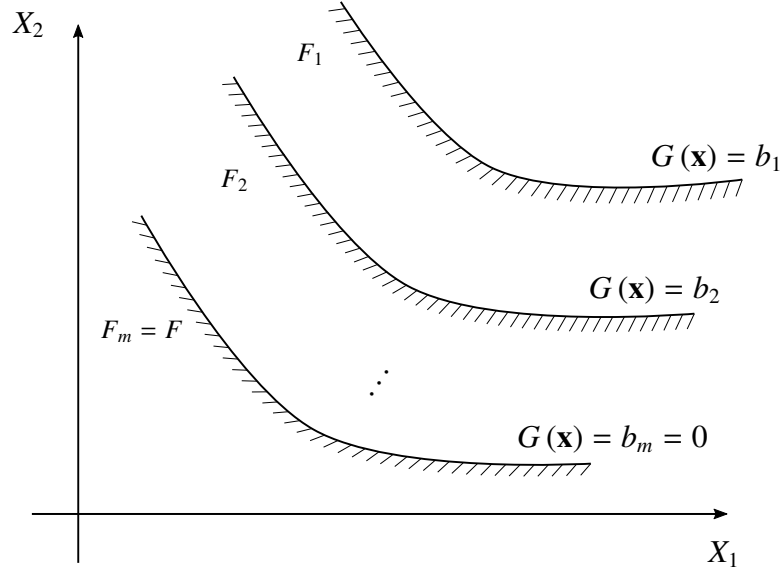
Under these circumstances, the only practical solution to accurately approximate equation 2-5 is through simulation methods. Particularly, Monte Carlo simulation has proven to be robust to the dimension of the problem and the complexity of the limit state function. However, the main drawback of MCS is its efficiency, especially when computing small probabilities of failure (e.g.,  $P_f \leq 10^{-2}$ ). MCS needs a large number of simulations to ensure an acceptable accuracy in small probabilities of failure estimates, and this is computationally expensive. For this reason, alternative computations methods to avoid costly MCS have appeared, such as the *subset simulation algorithm* proposed by [Au and Beck \(2001\)](#).

The present chapter is devoted to elucidate the concepts and theory behind this procedure, and its role when computing small probabilities of failure.

### 3.1 Basic idea of subset simulation

The subset simulation algorithm, abbreviated as SubSim or SS, is an adaptive stochastic simulation that efficiently computes small probabilities of failure, but which in turn conserves robustness of MCS ([Au et al., 2007](#)). In this approach, the original problem of computing a small probability of failure is expressed as the product of larger values of conditional probabilities of intermediate failure events  $b_j$ . In other words, the original problem related to a rare event is converted into a sequence of simulations involving more frequent events.

Let  $F$  denote the failure event, and let  $F_1 \supset F_2 \supset \dots \supset F_m = F$  be a decreasing nested sequence of failure events, such that  $F_k = \bigcap_{j=1}^k F_j$  for  $k = 1, 2, \dots, m$ . This sequence is schematically represented in figure 3-1.



**Figure 3-1:** Scheme of a progressive sequence of failure regions ( $b_1 > b_2 > \dots > b_m = 0$ ) for the subset simulation method, considering two random variables  $X_1$  and  $X_2$  (adapted from [Uribe, 2011](#)).

Now, by definition of conditional probability, the probability of  $F$  can be expressed as:

$$\begin{aligned}
 P_f &= P(F_m) = P\left(\bigcap_{j=1}^m F_j\right) \\
 &= P\left(\bigcap_{j=1}^{m-1} F_j\right) P\left(F_m \mid \bigcap_{j=1}^{m-1} F_j\right) \\
 &= P(F_{m-1}) P(F_m | F_{m-1}) = \dots \\
 &= P(F_1) \prod_{j=2}^m P(F_j | F_{j-1})
 \end{aligned} \tag{3-1}$$

Equation (3-1) states that probability of failure  $F$  can be expressed as a product of a sequence of conditional probabilities  $\{P(F_j | F_{j-1}) : j = 2, 3, \dots, m\}$  and  $P(F_1)$ . In other words, the probability of failure is equal to the probability of the final subset  $F_m$  given that in the previous step the event belongs to the subset  $F_{m-1}$ , but in turn the probability of  $F_{m-1}$  is conditioned by the event  $F_{m-2}$ , and so sequentially until  $F_1$  ([Au and Beck, 2001](#)).

In this way, even if a probability of failure is too small, by choosing the intermediate failure events  $\{F_j : j = 2, 3, \dots, m\}$  conveniently, the conditional probabilities of equation (3-1) can be made sufficiently large to be efficiently computed through simulation methods. Thus, the initial problem of simulating a rare event in the original probability space becomes a sequence of more frequent events in the conditional probability space which are much easier to simulate.

For estimating  $P_f$  through equation (3-1) is necessary to compute the probabilities  $P(F_1)$  and  $\{P(F_j|F_{j-1}) : j = 2, 3, \dots, m\}$ . The former can be easily estimated by the crude MCS:

$$P(F_1) \approx \frac{1}{N} \sum_{k=1}^N \mathbb{I}_{F_1}[\mathbf{x}_k \in F_1] \quad (3-2)$$

where  $\{\mathbf{x}_k : k = 1, 2, \dots, N\}$  are independent and identically distributed (i.i.d.) samples drawn from the joint PDF  $f_X(\mathbf{x})$ .

The conditional probabilities  $P(F_j|F_{j-1})$  are computed based on a estimator similar to equation (3-2). However, in this case, the samples must be simulated according to the conditional distribution of  $X$  given that it lies in  $F_{j-1}$ . This task could be carried out employing MCS, but it would be inefficient, and the simulation would fall into the same original problem. Fortunately, this task can be efficiently done by employing the Markov chain Monte Carlo MCMC. Particularly, a customized version of the Metropolis-Hastings algorithm is used for this purpose.

The following sections proceed as follows: Section 3.2 gives an overview of Markov Chain Monte Carlo (MCMC) methods, with an emphasis on the Metropolis and Metropolis-Hastings algorithms; on the other hand, the bulk of the subset simulation algorithm is explained in section 3.3; finally, the application of SubSim in geotechnical engineering is presented in section 3.4 by means of two practical examples.

## 3.2 Markov-Chain Monte Carlo methods

Markov Chain Monte Carlo (MCMC) methods comprise a series of algorithms aimed to draw samples from complex probability distributions, usually multidimensional, that cannot be directly sampled, at least not in an efficient way. Thus, these algorithms are especially useful for evaluating multi-dimensional integrals or constructing histograms of complex probability distributions.

Essentially, MCMC is a Monte Carlo simulation that relies on Markov Chains. These methods are based on constructing a Markov chain whose stationary distribution coincides with the probability distribution of our interest (Zuev et al., 2012). Thus, while MCS draws samples from the desired distribution for obtaining sample averages in order to approximate expectations, MCMC methods draw these samples from the long time run of an appropriated Markov Chain whose stationary distribution is equivalent to the distribution of our interest (Gilks et al., 1995). Nonetheless, whereas random samples obtained through MCS are statistically independent, samples drawn through MCMC methods are auto-correlated, which is an undesirable property that must be tackled.

There are several MCMC algorithms available in the literature. However, this document is focused in the Metropolis and the Metropolis-Hastings algorithms. The concepts of these two algorithms will serve to understand the Modified Metropolis algorithm (MMA) employed by the SubSim

procedure. More details of MCMC methods, and their algorithms, can be found in specialized references, such as [Gilks et al. \(1995\)](#) or [Walsh \(2004\)](#).

### 3.2.1 Metropolis algorithm

The Metropolis algorithm was proposed by [Metropolis et al. \(1953\)](#) for the specific case of approximating the Boltzmann distribution, but given its generality, it was posteriorly applied to other probability distributions. This procedure is considered the first MCMC method and receives its name after its main author, Nicholas Metropolis. It is specially useful for approximating and computing integrals of multi-dimensional distributions.

The Metropolis algorithm employs an ergodic Markov Chain that meets the reversibility property (see definition 2.1.6) with respect to the desired distribution. Its power lies in the ability to draw samples from the target distribution  $\pi(\mathbf{x})$  by only knowing its shape, or a function proportional to its shape. Thus, Metropolis algorithm avoids the computation of the normalization factor, which is usually extremely tedious in practice.

When employing the Metropolis algorithm, samples are generated iteratively following a Markov chain structure, such that the distribution of the next sample  $\mathbf{x}_{i+1}$  depends only on the current sample  $\mathbf{x}_i$ , and not in the previous ones. This procedure is performed through a proposal distribution  $q(\cdot|\mathbf{x}_i)$ , symmetric per definition (e.g., normal or uniform distributions), that proposes a new sample  $\boldsymbol{\xi}$  based on the current  $\mathbf{x}_i$ . However, not all generated samples are accepted, but rather are accepted or rejected depending of a comparison. The candidate sample is accepted when a random number  $u$  drawn from the standard uniform distribution  $\mathcal{U}(0, 1)$  is less or equal to the following value:

$$\alpha(\mathbf{x}_i, \boldsymbol{\xi}) = \min\left(\frac{\pi(\boldsymbol{\xi})}{\pi(\mathbf{x}_i)}, 1\right) \quad (3-3)$$

otherwise the candidate sample is rejected.

If the candidate sample  $\boldsymbol{\xi}$  is accepted, then  $\mathbf{x}_{i+1} = \boldsymbol{\xi}$ ; otherwise, the next sample is equal to the current one  $\mathbf{x}_{i+1} = \mathbf{x}_i$ . In this way, sometimes the chain will move from one state to another, and sometimes it will remain in the same state. Note that when the proposed sample  $\boldsymbol{\xi}$  is of more likelihood than the current, it will be accepted. Nonetheless, when the proposed sample is of less likelihood than the current, it can be either accepted or rejected. The less the likelihood of  $\boldsymbol{\xi}$ , the more likely to reject the proposed sample. Therefore, the chain will tend to stay in the high-density regions of  $\pi(\mathbf{x})$ , while occasionally it will visit low density regions of the target distribution.

For a better understanding of the methodology, algorithm 1 summarizes the procedure for drawing samples from a probability distribution by employing the Metropolis algorithm. This algorithm will be also applicable to the Metropolis-Hastings procedure that will be introduced in the next section.

---

**Algorithm 1** Metropolis and Metropolis-Hastings Algorithm (adapted from [Uribe, 2011](#))

---

**Input:**

- 1: ▶ Initial state of the chain  $\mathbf{x}_1$
- 2: ▶ Total number of states-samples  $N$
- 3: ▶ Target distribution  $\pi$
- 4: ▶ Proposal distribution  $q(\cdot|\mathbf{x}_i)$

**Algorithm:**

- 5: Initialize  $\mathbf{x}_1$
- 6: **for**  $i = 1$  to  $N$  **do**
- 7:     Sample  $u \sim \mathcal{U}(0, 1)$
- 8:     Sample  $\xi \sim q(\cdot|\mathbf{x}_i)$
- 9:     **if**  $u \leq \alpha(\mathbf{x}_i, \xi)$  **then**
- 10:          $\mathbf{x}_{i+1} = \xi$
- 11:     **else**
- 12:          $\mathbf{x}_{i+1} = \mathbf{x}_i$
- 13:     **end if**
- 14: **end for**
- 15: **return**  $\mathbf{x}$

**Output:**

- 16: ▶  $\mathbf{x}$  states of the Markov chain with stationary distribution  $\pi$
- 

Despite the robustness and the advantages of the Metropolis Algorithm, this methodology has several disadvantages or rather considerations that must be taken into account. In the first instance, proposal distributions are restricted to symmetrical distributions, which in some cases may be inconvenient if an asymmetric distribution performs better. On the other hand, drawn samples suffer from autocorrelation, which is an undesirable property that must be avoided. Furthermore, the initial value of the chain has a strong incidence on the first section of the chain, i.e. all the early values before the chain approaches its stationary distribution. Finally, this algorithm is not recommended for problems in very high-dimensions because of the slow convergence, since for larger the dimensions, the acceptance probability  $\alpha$  becomes very small, and hence many repeated samples appear ([Katafygiotis and Zuev, 2008](#)).

Fortunately, all the above issues can be solved by implementing different actions. For example, the autocorrelation may be mitigated by using a thinning or *lag period*; whereas the issue of the chain initial value is solved by implementing the *burn-in* period, i.e. by discarding the early values obtained before the chain converges to its stationary distribution. These actions carry a computational cost, but in contrast to the benefits of the algorithm this is only a trifle. Additionally, Metropolis algorithm can be extended to implement asymmetric proposal distributions, in the so-called *Metropolis-Hastings algorithm*. The above concepts will be discussed in the next section when explaining the Metropolis-Hastings algorithm. Finally, [Au and Beck \(2001\)](#) tackled the issue

of convergence in high-dimensions, developing thus the *Modified Metropolis-Hastings algorithm* that will be described in section 3.3 for order purposes.

Finally, it is worth presenting a toy example to consolidate all these concepts. In the following exercise, none of the above actions to improve sampling will be implemented, so the best Markov chain is not expected. Later, after explaining in more detail the actions to improve Markov chains, a new example will be presented in which all these techniques are applied. A comparison of these two examples will highlight the importance of implementing different actions to improve sampling.

**Example 3.2.1.** Let us consider the following probability distribution:

$$\pi(x) = \begin{cases} \frac{\sin(x) + 1}{\int_{-\frac{5\pi}{2}}^{\frac{3\pi}{2}} \sin(x) + 1 \, dx}, & -\frac{5\pi}{2} \leq x \leq \frac{3\pi}{2} \\ 0, & \text{otherwise} \end{cases}$$

from which it is required to obtain  $N = 1000$  samples, plot the respective histogram in contrast with the theoretical function, and plot the Markov Chain trajectory.

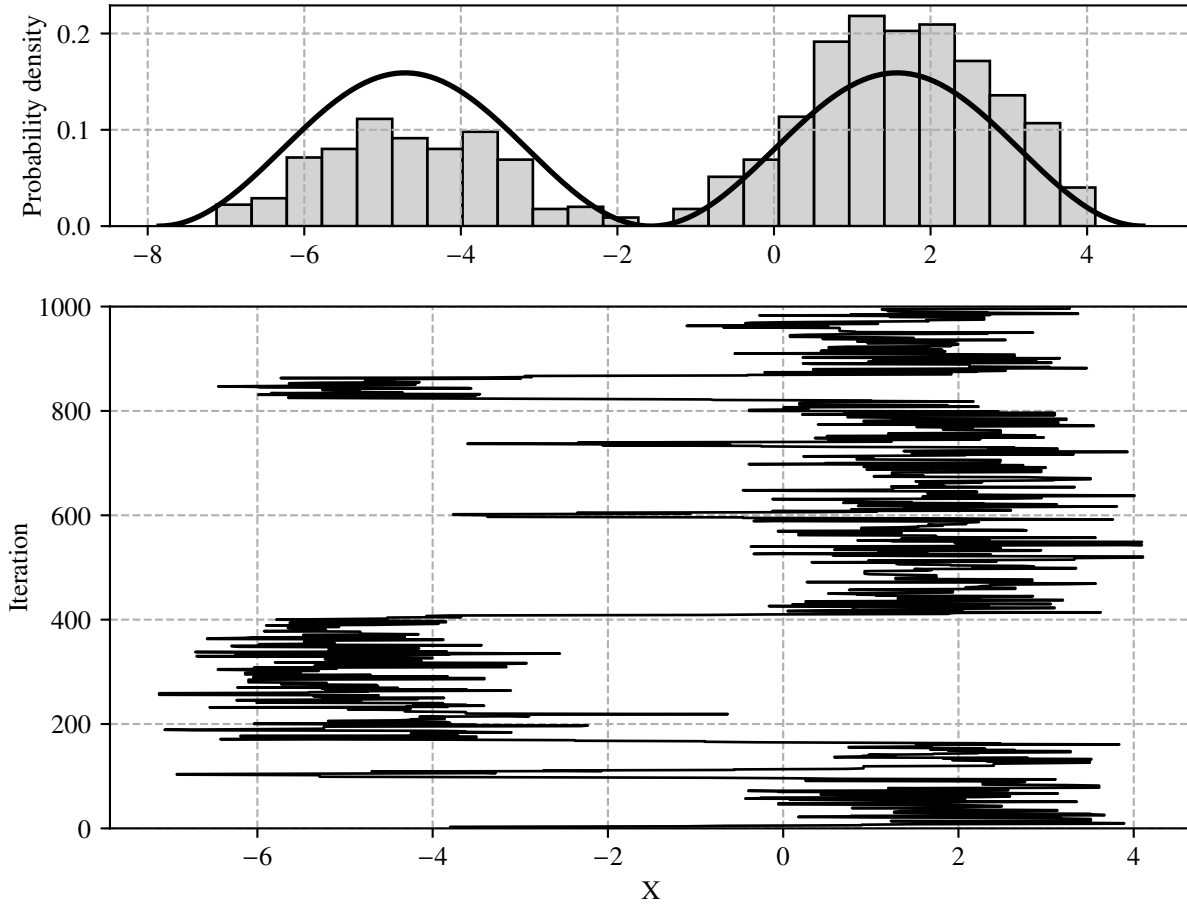
Note that by virtue of the Metropolis algorithm, the scaling factor of  $\pi(x)$ , i.e. the integral on the denominator, does not need to be computed. Instead, only the shape of the distribution is required, that is  $\pi(x) \propto \sin(x) + 1$ .

To solve this problem, the procedure described by the algorithm 1 is followed. The proposal distribution  $q(\cdot|x_i)$  is taken as a uniform distribution centered at  $x_i$ , with a dispersion equal to 3, i.e.  $q(x_{i+1}|x_i) \sim \mathcal{U}(x_i - 3, x_i + 3)$ . Furthermore, the chain initial state is defined as  $x_1 = 0$ , and a total number of  $N = 1000$  samples are drawn from the procedure.

After applying the Metropolis algorithm with these considerations,  $N = 1000$  samples are obtained from  $\pi(x)$ , with an acceptance rate of 0.554. The histogram and the chain trajectory of the samples are presented in figure 3-2. It is evident from figure 3-2 that the chain has long flat periods (i.e., it is *poorly mixing*), and that the histogram does not perfectly match the theoretical distribution. These drawbacks are due to not implementing the burn-in period and a not sufficient number of iterations to achieve the stationary distribution, furthermore, the spread of the proposal distribution may not be the most appropriate.

### 3.2.2 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is an extension, proposed by Hastings (1970), of the Metropolis algorithm. Both algorithms have the same concept and follow a similar procedure, but the Metropolis-Hastings algorithm has the advantage of including asymmetric distributions as an option of proposal distributions.



**Figure 3-2:** Sampling a probability distribution employing the Metropolis Algorithm. Top: normalized histogram and target distribution  $\pi(x)$ . Bottom: Markov chain trajectory.

In this way, by employing the Metropolis-Hastings algorithm the next sample  $x_{i+1}$  is defined based on the proposed sample  $\xi$ , which is obtained through a transitional distribution  $q(\xi|x_i) = P(x_i \rightarrow \xi)$ . In this case, the proposal distribution does not need to be symmetrical, so the range of options is expanded and the sampling from the target distributions could be better and faster. Finally, the candidate sample  $\xi$  is accepted when a random number  $u$  drawn from the standard uniform distribution  $\mathcal{U}(0, 1)$  is less or equal to:

$$\alpha(x_i, \xi) = \min\left(\frac{\pi(\xi)q(x_i|\xi)}{\pi(x_i)q(\xi|x_i)}, 1\right) \quad (3-4)$$

and rejected otherwise.

Evidently, Metropolis-Hastings algorithm is the same Metropolis algorithm but including asymmetric distributions and implementing equation (3-4) instead of equation (3-3). Therefore, algorithm 1 also holds for the Metropolis-Hastings case, and only the way of computing  $\alpha(x_i, \xi)$ , in line

9, changes. Note that when using a symmetrical proposal distribution in the Metropolis-Hastings algorithm, equation (3-4) is equivalent to equation (3-3).

Nonetheless, although the Metropolis-Hastings algorithm allows to include asymmetric proposal distributions in the procedure, it has the same other disadvantages of the original Metropolis algorithm. Again, random samples are auto-correlated, and the starting value has a strong incidence in the first section of the chain. Fortunately, these pitfalls can be solved through several techniques and recommendations, some of which will be described below.

### 3.2.3 Techniques to improve Markov Chains

This section explains some techniques for improving sampling when using the Metropolis and the Metropolis-Hastings algorithms.

#### 3.2.3.1 Chain starting value

The stationary distribution of the Markov chain is not affected by its initial value  $\mathbf{x}_i$  when the chain is irreducible (definition 2.1.4). Nonetheless, an inappropriate initial value may lead to a lengthy burn-in period, that is many samples have to be obtained before the Markov chain approaches its stationary distribution. If the proposal distribution is suitable for the Markov chain (*well-mixing*), that is the Markov chain has a good initial distribution in the sampling distribution, it will quickly approach its stationary distribution regardless of its starting value. However, if the sampling distribution is not adequate for the chain (not well-mixing), then the burn-in period will be lengthy if extreme initial values are employed. A good practice would be to start the chain the closest to the center of the target distribution, e.g. at the mode of the distribution (Walsh, 2004).

#### 3.2.3.2 Definition of the proposal distribution

As stated before, the choice of the proposal distribution  $q(\cdot|\mathbf{x}_i)$  will directly affect the rate of convergence of the chain to its stationary distribution  $\pi^*$  (Gilks et al., 1995). In this sense, by adjusting the proposal distribution parameters, the sampling of the target distribution will be improved. To this end, the suitable parameters of the proposal can be defined by examining the relationship between the acceptance probability and the  $k$ -order auto-correlations  $R(k)$  of the samples. This relationship can be plotted to show the variation of the auto-correlation for different values of acceptance probability. In this way, the optimal parameters of the proposal distribution can be defined as those for which the auto-correlation is minimum. Nonetheless, optimal parameters obtained through this procedure must be treated carefully, since they only guarantee a decrease of the auto-correlation but not a faster convergence of the chain. For this latter purpose, it is necessary to employ other techniques to improve sampling.

### 3.2.3.3 Length of the chain

There exist several opinions about the suitable Markov chain length, some of them contradictory (Gilks et al., 1995). Nonetheless, as a general rule, a very long run of one or several chains will increase the probability of populating (sampling) the entire region of the target distribution. In essence, the Markov chain length mostly depends on the stage of burning, the required amount of random samples, and the dimension of the target distribution.

### 3.2.3.4 Burn-in period

As stated before, the burn-in period is the number of samples until the Markov chain approaches to its stationary distribution. The definition of this period is a fundamental aspect when performing a Metropolis or Metropolis-Hastings algorithm. It depends on the chain starting value  $x_1$ , on the rate of convergence to stationary distribution, on the desired similarity between the simulated chain and the stationary distribution, and on the suitability of the proposal distribution (Gilks et al., 1995). All these aspects may increase the length of the burn-in period, and therefore have an impact on the algorithm efficiency.

In the literature, there exist several alternatives to estimate the burn-in period length. The most common method to identify the burn-in period is by plotting the chain trajectory to inspect the number of samples until the Markov chain becomes stationary. Another interesting alternative to estimate the burn-in period was employed by Uribe (2011). In this latter approach, the auto-correlation of the initial samples is computed, and when it starts to decrease the corresponding number of samples is set as the burn-in period. Furthermore, other authors have argued that the computation of the burn-in period length is unnecessary since this is likely to be equal or less than 1% or 2% of the total length of a Markov chain sufficiently long (Gilks et al., 1995); typically, this number corresponds to the first 1000 to 5000 samples.

### 3.2.3.5 Thinning or lag period

In some cases, the Metropolis, or the Metropolis-Hastings algorithm, draws highly auto-correlated samples from the target distribution. This is especially evident when the proposal distribution is not the appropriated one, and hence the rate of acceptance is slow. Under this situation, the Markov chain gets stuck in one location for long periods, corresponding to all the rejected  $\xi$ . To avoid this undesirable property, one strategy is to implement the thinning or lag periods. This strategy consists in carrying out multiple iterations of the algorithm, after the burn-in period, and between successive samples each lag-th position is stored and the rest are discarded. Note that implementing the lag period saves computer memory but increases the chain length for a required number of random samples (Walsh, 2004).

### 3.2.3.6 Test for convergence

The tests for convergence are more formal diagnostics to assess whether the Markov chain has reached its stationary distribution, or in other words, that the burn-in period has been well-defined. There exist several tests of this type in the literature, but this is actually an active area of study that needs a lot of research, especially in the case of high dimensions or when the rate of convergence is slow (Gilks et al., 1995). Bearing these considerations in mind, the utilization of the Geweke test is proposed in this work.

The Geweke test consists in comparing the first 10% and the last 50% sections of the chain, after the burn-in period. If the chain is stationary, the respective means of these two sections are approximately equal; otherwise, the chain has not reached its stationary distribution, and hence the burn-in period is not well applied and then it must be redefined (Walsh, 2004).

The following example illustrates the improvements achieved by implementing the aforementioned strategies when conducting a Metropolis or Metropolis-Hastings algorithm.

**Example 3.2.2.** (adapted from Stansbury, Dustin (2012)) Let us consider the following probability density function:

$$\pi(x) = \begin{cases} \frac{\frac{A^x}{\Gamma(x)} B^{x-1} \exp^{-AB} \sin(x\pi_c)}{\int_0^{\infty} \frac{A^x}{\Gamma(x)} B^{x-1} \exp^{-AB} \sin(x\pi_c) dx}, & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

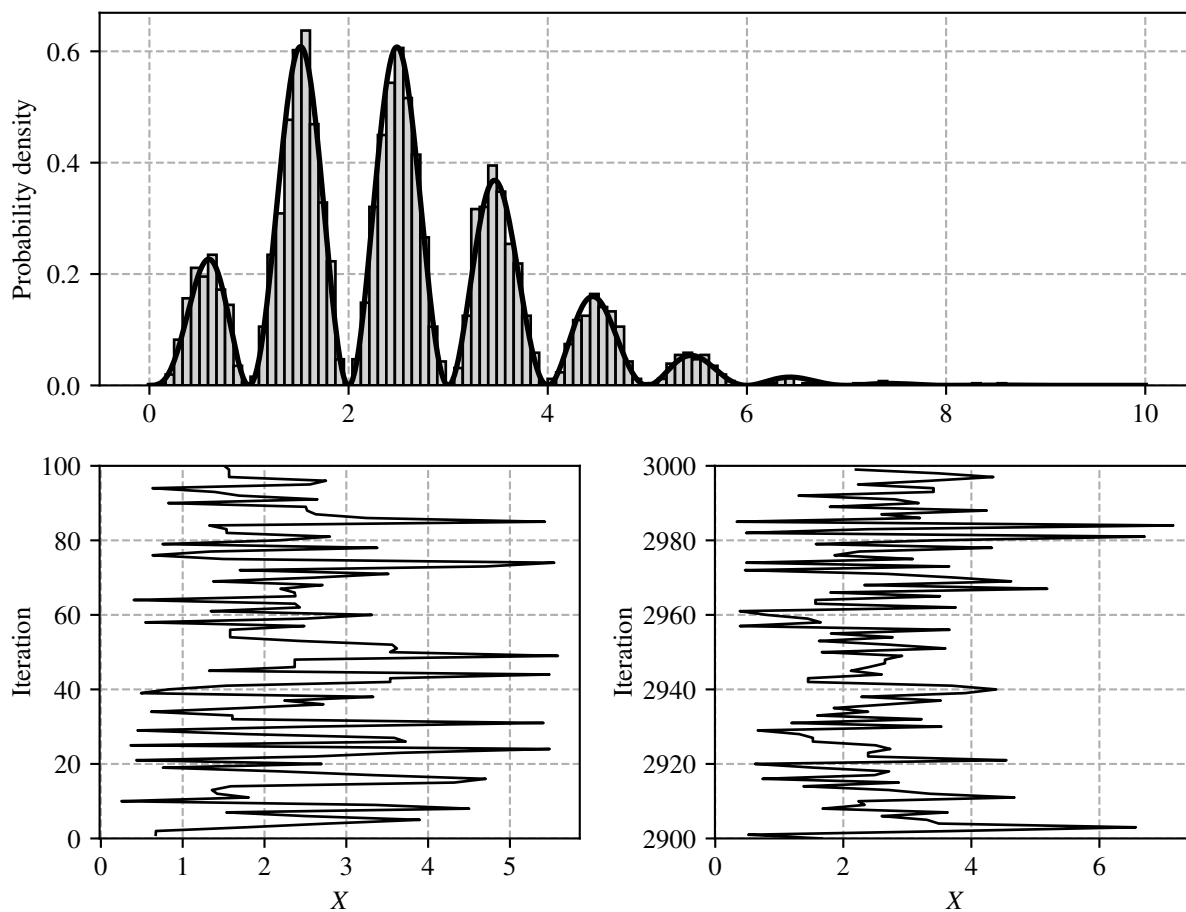
where  $\Gamma$  is the Gamma function,  $A = 1.0$  and  $B = 1.5$  are adjustment parameters, and  $\pi_c$  is the ubiquitous mathematical constant equal to  $3.141592\dots$  and must not be confused with the target distribution  $\pi$ .

It is required to obtain  $N = 3000$  samples from  $\pi(x)$  using the Metropolis or Metropolis-Hastings algorithm. Additionally, it is needed to plot the histogram of the samples and compare it with the theoretical function, plot the Markov chain trajectory, and plot the correlogram of the first and last 10% sections of the chain. Techniques for a better sampling previously stated must be employed, and the Geweke test has to be performed to validate the chain stationarity. Note again that, by virtue of the Metropolis and the Metropolis-Hastings algorithms, the scaling factor of  $\pi(x)$  (that is the integral on the denominator) does not need to be computed, and only the shape of the distribution is required, that is,  $\pi(x) \propto \frac{A^x}{\Gamma(x)} B^{x-1} \exp^{-AB} \sin(x\pi_c)$ .

The next step is to define the chain starting value  $x_1$  and the proposal distribution. After examining the target function,  $x_1 = 3$  is defined as a good starting value since it is approximately in the middle of the function. On the other hand, a normal distribution centered in the current sample is selected

as the proposal distribution. Additionally, several standard deviations were evaluated to improve the sampling, and it was defined that  $\sigma = 4$  is a good value. This examination was done visually employing several standard deviations; however, all these graphs are not presented here for not saturating the document. Thus, the proposal distribution is defined as  $q(\xi|\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i, 4)$ , where  $\mathcal{N}$  stands for the normal distribution.

Furthermore, a lag period equal to 10 will be employed to reduce samples auto-correlation, and a burn-in period of 2% of the total length of the chain will be adopted, hoping that all the samples are drawn from the stationary distribution. The autocorrelation and the stationarity will be corroborated after the sampling through the correlograms and the Geweke test, respectively. Note that for obtaining 3000 samples, it is required to first draw 30600 samples from the algorithm, in order to throw out the first 600 samples (burn-in period) and apply the lag period of 10.



**Figure 3-3:** Samples drawn in example 3.2.2 through the Metropolis-Hastings algorithm using techniques for better Markov chains. Top: normalized histogram of the samples and target distribution. Bottom left: Trajectory of the first 100 samples of the chain. Bottom right: Trajectory of the last 100 samples of the chain

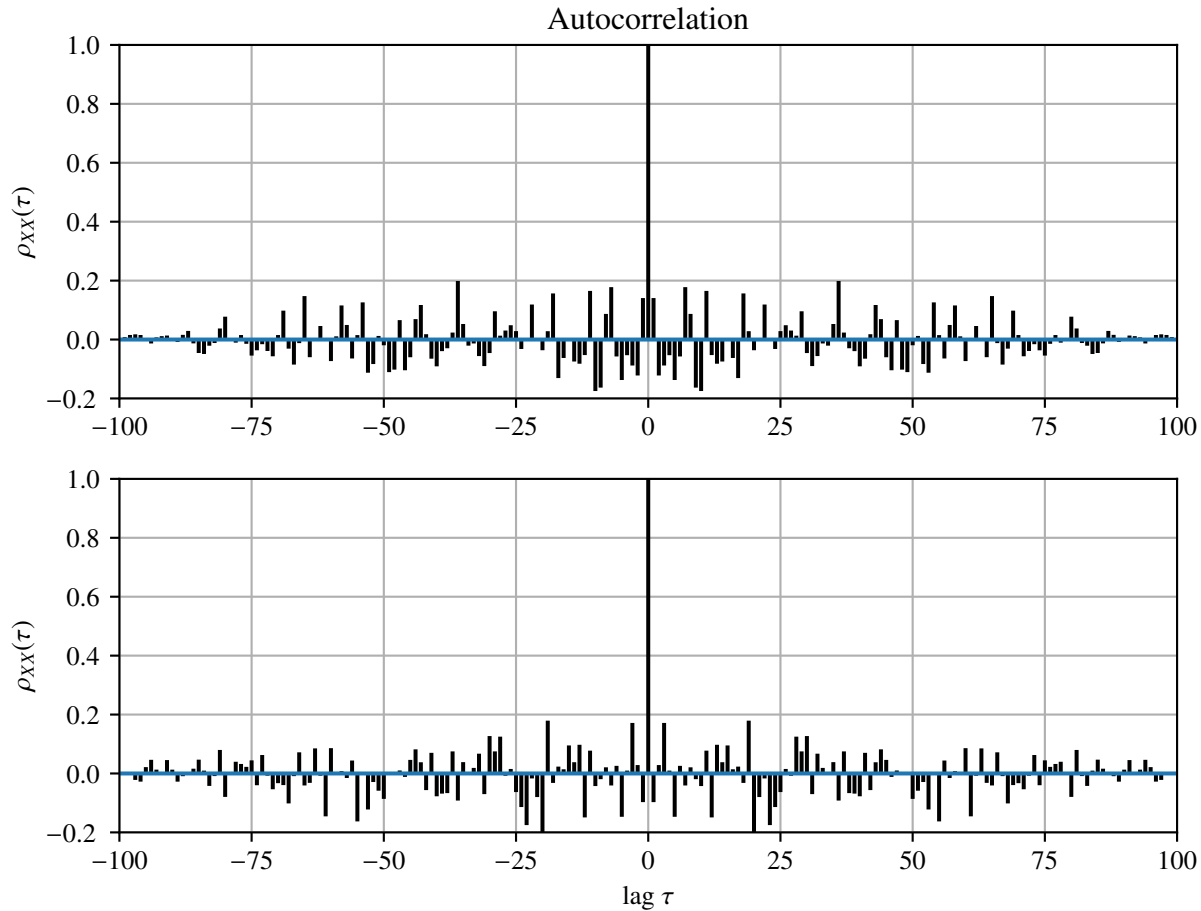
Bearing these considerations in mind, algorithm 1 is applied to obtain the required random samples from the target distribution  $\pi(x)$ . The resulting samples are represented by the histogram in figure 3-3, where the theoretical curve is also plotted. Additionally, figure 3-3 also plots the Markov chain trajectory for the first 100 samples and the last 100 samples.

From figure 3-3 it is evident that the shape histogram fits very well to the target distribution. The chain trajectory indicates that the chain is well-mixed, and flat periods are negligible. At first glance, it could be said that the chain reached its stationary distribution and that the samples are obtained according to the target distribution. These results contrast with those obtained in the example 3.2.1 and presented in figure 3-2, where the histogram of the drawn samples does not correctly fit the theoretical curve. By comparing figure 3-3 and figure 3-2, it is evident the superiority of the sampling when techniques for better Markov chains are implemented, demonstrating thus its importance.

Now, the other issue corresponds to the auto-correlation of the obtained samples. To analyze this property the correlograms of the first and last 100 samples are plotted in figure 3-4. From this figure is evident the low rate of auto-correlation of the analyzed samples. The correlogram of the first 100 samples shows that these are rather independent, which indicates that the burn-in period is well defined and that the chain has already reached its stationary distribution. On the other hand, the correlogram of the last 100 samples also shows that these samples are rather independent, which indicates stationarity and that the chain length is adequate. So, the auto-correlation is satisfactorily mitigated by implementing the burn-in and lag periods.

Furthermore, it is found that the acceptance rate of this example is equal to 0.24. This value indicates that 24% of the time the proposed samples are accepted, and hence the other 76% are rejected. This acceptance rate is in concordance with the values recommended by the literature, which oscillate close to an acceptance rate between 23%-25% (see Gelman et al., 1996; Zuev et al., 2012).

Finally, the Geweke test is performed to corroborate the suitability of the chain. In this regard, the mean of the first 10% samples of the chain is estimated equal to 2.47448 and the mean of the last 50% samples of the chain is estimated equal to 2.472166, so the suitability of the chain is approved in light of the Geweke test (< 3% error). This result shows that the chain reached its stationary distribution and that the drawn samples correctly follow the target distribution. Note that this example and example 3.2.1 employ algorithm 1; however, the sampling is drastically improved by implementing the above techniques for a better Markov chain. These techniques will carry an additional computational cost, but in contrast to the advantages of the algorithm and the improvements achieved, this computational cost is only a trifle.



**Figure 3-4:** Correlograms of the samples obtained in figure 3-3. Top: Autocorrelation between the first 100 samples of the chain. Bottom: Autocorrelation between the last 100 samples of the chain

### 3.3 Subset simulation

As stated before, subset simulation express the probability of failure of a geotechnical model, or any mechanical system, through equation (3-1). For solving this equation, it is needed to define a decreasing nested sequence of failure events  $F_1 \supset F_2 \supset \dots \supset F_m = F$ , obtain the probability of occurrence of the first subset ( $P(F_1)$ ) and the subsequent sequence of conditional probabilities for each subset  $\{P(F_j|F_{j-1}) : j = 2, 3, \dots, m\}$ .

The probability of occurrence of the first failure event  $p_1 = P(F_1|F_0) = P(F_1)$  is straightforward to approximate by a crude MCS, that is equation (3-2), since samples are directly drawn from the joint PDF  $\pi$  of the model. However, two main issues appear when computing the conditional probabilities  $\{P(F_j|F_{j-1}) : j = 2, 3, \dots, m\}$ . The first one concerns to the choice of the intermediate failure events. Most of the algorithm efficiency depends on this selection. The second issue deals

with the sampling from the conditional distributions  $\pi(\cdot|F_{j-1})$ , i.e. to only draw samples from  $\pi(\cdot)$  under the threshold  $b_{j-1}$  that defines  $F_{j-1}$ , such that

$$F_{j-1} = \{\mathbf{x} \in \mathbb{R}^d : G(\mathbf{x}) > b_{j-1}\},$$

in order to obtain the conditional probabilities  $P(F_j|F_{j-1})$ .

This section is devoted to explain the subset simulation algorithm for computing equation (3-1), but for this purpose, the above two issues will be firstly addressed. In particular, it will be explained how intermediate failure events can be adaptively selected, and a modified version of the Metropolis-Hastings algorithm will be introduced for the task of drawing conditional samples.

### 3.3.1 Choice of intermediate failure events

The definition of the intermediate failure events  $\{F_j : j = 1, \dots, m\}$  plays an important role in the subset simulation algorithm. In this regard, it is necessary to consider two main issues. Firstly, the parametrization of the target failure domain  $F$ , which allows generating intermediate domains by varying a defined parameter. Secondly, the choice of the sequence of values of the defined parameter, which greatly affects the the conditional probabilities  $P(F_j|F_{j-1})$  and hence the efficiency in the convergence of the algorithm (Au and Beck, 2001).

Regarding the first issue, subset simulation employs the performance function  $G$  to parametrize the model domain. In this sense, the failure domain can be defined as the exceedance, through the performance function, of some prescribed critical threshold or limit capacity  $b$ :

$$F = \{\mathbf{x} \in \mathbb{R}^d : G(\mathbf{x}) > b\}$$

and therefore, the sequence of intermediate failure events can be expressed as:

$$F_j = \{\mathbf{x} \in \mathbb{R}^d : G(\mathbf{x}) > b_j\}$$

where  $b_1 > b_2 > \dots > b_m = b$  is an increasing sequence of threshold values that leads to the system failure  $F$  when the critical threshold  $b_m = b$  is exceeded.

The second issues deals with the selection of the sequence of intermediate thresholds, that is  $\{b_j : j = 1, \dots, m\}$ , which define the nested failure events. Note that if the sequence of intermediate thresholds increases slowly (i.e.  $b_j \approx b_{j-1}$ ), then the conditional probabilities will be larger and, in principle, would require fewer samples  $N$  for their computation. Nonetheless, a slow sequence will require more levels of simulation ( $m$ ) to reach the final failure domain, increasing thus the total number of samples in the whole procedure ( $N_T = mN$ ). Conversely, if the sequence of intermediate thresholds increases too quickly (i.e.  $b_j \gg b_{j-1}$ ), then the conditional failure events become rarer and hence their probability smaller, increasing thus the number of samples per each intermediate level  $N$  and the total number of samples in the whole procedure  $N_T$ . Therefore, the choice of the

intermediate threshold is a trade-off between the number of samples required in each level and the number of levels required to reach the final failure domain (Au et al., 2007).

In most practical applications, it is quite difficult to define the optimal intermediate threshold values in order to obtain reasonable estimates of the conditional probabilities. In other words, it is difficult to control conditional probabilities by choosing intermediate threshold values a priori. An alternative is then to define  $b_j$  values adaptively, based on a prescribed conditional probability  $p_0 \in (0, 1)$ , common to all intermediate levels. In this latter approach, the intermediate threshold value  $b_j$  is defined as the  $[(1 - p_0)N]$ -th largest value among  $\{G(\mathbf{x}_j^{(i)}) : i = 1, \dots, N\}$  where  $\mathbf{x}_j^i$  are the Markov chain samples obtained at the  $j$ -th conditional level for  $j = 1, \dots, m$ , and  $\mathbf{x}_0^i$  are the samples obtained in the initial MCS. In this way, the choice of the intermediate threshold values will depend on the prescribed conditional probability  $p_0$ , and hence on the conditional samples, so  $b_j$  will vary in different simulation runs.

Much has been discussed in the literature about the optimal value of  $p_0$  for improving the algorithm efficiency. Nonetheless, Zuev et al. (2012) demonstrated that choosing  $p_0 \in [0.1, 0.3]$  will lead to a similar efficiency as that obtained by the optimal value of  $p_0$ , so there is no need to refine its definition as long as subset simulation is implemented properly. Furthermore, it is desirable to set a  $p_0$  value so that  $p_0N$  and hence  $(1 - p_0)N$  are positive integers, although it is not strictly mandatory. Finally, it is worth noting that  $P(F_j|F_{j-1})$  is not actually equal to  $p_0$ , since it is an estimate of the conditional probability that depends on the number of samples  $N$  in each level. Thus,  $N$  must be large enough so that the error in the approximation  $P(F_j|F_{j-1}) \approx p_0$  can be guaranteed.

### 3.3.2 Modified Metropolis algorithm

Subset simulation algorithm requires the generation of independent samples according to the conditional distribution  $\pi(\cdot|F_{j-1})$  in order to populate the intermediate failure regions and reach the failure domain. The MCMC methods, in particular, the Metropolis and the Metropolis-Hastings algorithms, were previously introduced as useful tools to simulate samples from conditional or complex probability distributions. At first sight, any of these two algorithms could be employed; however, there are two considerations to take into account for their implementation. Firstly, the efficiency of the MA and MHA is reduced in high-dimensions because of the low rate of convergence of the chain, which may lead to ergodicity problems. Secondly, simulated samples must lie below the intermediate threshold  $b_{j-1}$ . To overcome these two issues, Au and Beck (2001) proposed a new version of the original Metropolis algorithm, especially useful for the procedure of SubSim, and called it the *modified Metropolis algorithm* (MMA).

MMA differs from the original MA (and MHA) in two main aspects: first, unlike MA, that uses a multivariate proposal PDF on  $\mathbb{R}^d$  to obtain a candidate state  $\boldsymbol{\xi} = [\xi_1, \xi_2, \dots, \xi_k, \dots, \xi_d]$ , MMA employs a sequence of univariate proposal PDFs. In other words, in a MMA each coordinate  $\xi_k$

of the candidate state is obtained separately using a univariate proposal distribution conditioned to the current state  $q_k(\cdot|x_k)$ . Second, MMA has an additional step to the original MA, in which it checks whether the candidate state lies in the intermediate failure region  $F_{j-1}$ . If so, the candidate state is accepted; otherwise, the candidate state is rejected, and the current MCMC is repeated.

The MMA proceeds as follows: let  $q_k(\xi_k|x_k)$ , with  $k = 1, \dots, d$ , be a one-dimensional proposal PDF for  $\xi_k$  centered at  $x_k$  and with the symmetry property  $q_k(\xi_k|x_k) = q_k(x_k|\xi_k)$ , then:

1. *Generate a candidate state  $\xi_k$* : For each element  $k = 1, \dots, d$  generate a precandidate  $\tilde{\xi}_k \sim q_k(\cdot|x_k^{(i)})$ , and compute the acceptance ratio:

$$r = \frac{\pi_k(\tilde{\xi}_k)}{\pi_k(x_k^{(i)})} \quad (3-5)$$

and then set the candidate state  $\xi_k = \tilde{\xi}_k$  with a probability  $\min(1, r)$  or set  $\xi_k = x_k^{(i)}$  with a probability  $1 - \min(1, r)$

2. *Accept or reject  $\xi$* : If  $\xi \in F_j$  accepted it as the next sample,  $\mathbf{x}^{(i+1)} = \xi$ ; otherwise, reject  $\xi$  and take the current sample as the next sample,  $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)}$ .

Note that asymmetric univariate proposal distributions could also be implemented in this procedure, generating a generalized version that can be named as the modified Metropolis-Hastings algorithm (MMHA). In this case, equation (3-5) is rewritten in an analogous way to equation (3-4).

The first step of the MMA can be seen as a local random walk in the neighborhood of the current state  $\mathbf{x}_k^{(i)}$ , while the second step guarantees that the next sample lies in the intermediate failure domain  $F_{j-1}$ . Note that the proposal distributions play a key role in the procedure since these affect the deviation of the candidate state from the current state and hence impact the efficiency of the Markov chain. Nonetheless, it has been demonstrated that the type of proposal distributions do not affect the algorithm efficiency, but their spread does (Zuev et al., 2012). Therefore, there is no need to employ the MMHA since easily-handling symmetric proposal PDFs are sufficient. Small spreads increase the acceptance rate but also the dependence between successive samples due to their proximity, generating thus ergodicity problems, i.e. the failure regions may not be properly populated. Large spreads reduce the acceptance rate and hence increase the number of samples repeated, slowing down the convergence. Although the optimal choice of the proposal distribution spread depends on each particular case, in most cases a good choice is the uniform distribution centered at the current sample with a width equal to two times the standard deviation of the target distribution (Au and Beck, 2001). A more detailed discussion about this choice is given by Zuev et al. (2012).

To finalize this section it is worth highlighting some properties of the MMA. Firstly, given that the seed  $\mathbf{x}^{(1)} \sim \pi(\cdot|F_1)$ , then all the subsequent states  $\mathbf{x}^{(i)}$  will be automatically distributed according to the target distribution, i.e.  $\mathbf{x}^{(i)} \sim \pi(\cdot|F_j)$ , which leads to a *perfect sampling* property (see, e.g.,

Robert and Casella, 2013). Secondly, by virtue of the previous property, MMA does not need to implement the burn-in period since the seeds are already distributed according to the target distribution. Thirdly, a Markov chain starting at a single point may have ergodicity problems due to the existence of disconnected failure regions separated by safe regions whose size is large in contrast to the spread of the proposal PDF. However, this problem is solved by using several Markov chains starting at different points, as it does the MMA, since it is expected that different seeds lead to populate all the isolated failure regions. Algorithm 2 summarizes the MMA.

---

**Algorithm 2** Modified Metropolis algorithm (MMA) (adapted from Zuev et al., 2012)

---

**Input:**

- 1: ▶ Initial state of the chain  $\mathbf{x}_1$
- 2: ▶ Total number of states-samples ( $N$ )
- 3: ▶ Target distributions  $\pi_1(\cdot), \dots, \pi_d(\cdot)$ , i.e. marginal PDFs of  $x_1, \dots, x_d$
- 4: ▶ Univariate proposal PDFs depending on a current state  $q_1(\cdot|\mathbf{x}_{i,1}), \dots, q_d(\cdot|\mathbf{x}_{i,d})$

**Algorithm:**

- 5: **for**  $i = 1, \dots, N - 1$  **do**
- 6:     **for**  $k = 1, \dots, d$  **do**
- 7:         Sample  $u \sim \mathcal{U}(0, 1)$
- 8:         Sample  $\tilde{\xi}_k \sim q_k(\cdot|x_{i,k})$
- 9:         Compute  $\alpha(x_k^{(i)}, \tilde{\xi}_k) = \min\left(\frac{\pi(\tilde{\xi}_k)}{\pi(x_{i,k})}, 1\right)$
- 10:         **if**  $u \leq \alpha(x_k^{(i)}, \tilde{\xi}_k)$  **then**
- 11:             Set  $\xi_k = \tilde{\xi}_k$
- 12:         **else**
- 13:             Set  $\xi_k = x_k^{(i)}$
- 14:         **end if**
- 15:     **end for**
- 16:     **if**  $\xi \in F$  **then**
- 17:         Set  $\mathbf{x}_{i+1} = \xi$
- 18:     **else**
- 19:         Set  $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)}$
- 20:     **end if**
- 21: **end for**
- 22: **return**  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$

**Output:**

- 23: ▶  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$  states of the markov chain with stationary distribution  $\pi(\cdot|F)$
-

### 3.3.3 Estimation of the probability of failure

After discussing the two key issues of the SubSim procedure, i.e. the choice of the intermediate failure regions and the sampling through MMA, we can proceed to explain the main body of the algorithm. This section is devoted to this end, and the next section presents some practical examples.

As stated before, the subset simulation algorithm expresses the probability of failure of a system through the equation (3-1), so it is necessary to estimate the probability of occurrence of the first failure subset  $P(F_1)$  and the conditional probability of occurrence of the subsequent failure subsets  $P(F_j|F_{j-1})$ . Thus, it is required to first define the number of random variables involved in the problem  $d$ , the number of samples  $N$  per each conditional level, and the prescribed conditional probability  $p_0$ .

The unconditional probability of occurrence of the first failure subset  $P(F_1)$  is approximated by the crude MCS as:

$$P(F_1) \approx \hat{P}_1 = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\mathbf{x}_0^{(i)} \in F_1] \quad (3-6)$$

where  $\mathbf{x}_0^{(i)} : i = 1, \dots, N$  are i.i.d. samples drawn from the probability distributions of each parameter,  $\pi(\cdot)$ . In this case, the failure domain  $F_1$  is delimited by the first threshold value  $b_1$ , which is obtained as the  $p_0N$ -th position in the descending list of  $G(\mathbf{x}_0^{(i)})$  values.

Based on the initial  $p_0N$  samples that lie in  $F_1$ , the MMA is employed to generate  $[N(1 - p_0)]$  additional samples in order to obtain a total of  $N$  samples distributed according to  $\pi(\cdot|F_1)$  (Au and Beck, 2003). These samples are used to estimate  $P(F_2|F_1)$  by defining  $b_2$ , and hence  $F_2$ , in the same way as for the first failure subset, i.e. ordering the  $G(\mathbf{x}_1^{(i)})$  values. Samples that lie in  $F_2$  provide the seeds for simulating more samples, according to  $\pi(\cdot|F_2)$ , which are employed to estimate  $P(F_3|F_2)$  (Au and Beck, 2001). This process is repeated until the failure domain of interest  $F = F_m$  is reached. The probability of occurrence of these intermediate failure regions is estimated as:

$$P(F_j|F_{j-1}) \approx \hat{P}_j = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\mathbf{x}_{j-1}^{(i)} \in F_j] \quad (3-7)$$

Finally, by combining equations (3-6) and (3-7), the probability of failure can be expressed as:

$$\hat{P}_f = \prod_{j=1}^m \hat{P}_j$$

Note that the computational efficiency of the SubSim algorithm relative to the crude MCS increases with decreasing probability of failure (Au and Beck, 2003). Consequently, the true advantages of using the SubSim algorithm appear when computing small probabilities of failure. Algorithm 3 details the SubSim method.

---

**Algorithm 3** subset simulation algorithm (adapted from Zuev et al., 2012)

---

**Input:**

- 1: ▶ Conditional failure probability value,  $p_0$
- 2: ▶ Number of samples per conditional level,  $N$

**Algorithm:**

- 3: Set  $j = 0$
  - 4: Set  $N_F(j) = 0$ , where  $N_F(j)$  stands for the number of failure samples at level  $j$
  - 5: Sample  $x_0^{(1)}, \dots, x_0^{(N)} \sim \pi(\mathbf{x})$
  - 6: **for**  $i = 1, \dots, N$  **do**
  - 7:     **if**  $G^{(i)} = G(x_0^{(i)}) > b$  **then**
  - 8:          $N_F(j) \leftarrow N_F(j) + 1$
  - 9:     **end if**
  - 10: **end for**
  - 11: **while**  $N_F(j)/N < p_0$  **do**
  - 12:      $j \leftarrow j + 1$
  - 13:     Sort  $\{G^{(i)}\} : G^{(i_1)} \leq G^{(i_2)} \leq \dots \leq G^{(i_N)}$
  - 14:     Define  $b_j = \frac{G^{(i_{N-Np_0})} + G^{(i_{N-Np_0+1})}}{2}$
  - 15:     **for**  $k = 1, \dots, Np_0$  **do**
  - 16:         Starting from  $x_j^{(1),k} = x_{j-1}^{i_{N-Np_0+k}} \sim \pi(\cdot|F_j)$ , generate  $1/p_0$  states of a Markov chain  $x_j^{(1),k}, \dots, x_j^{(1/p_0),k} \sim \pi(\cdot|F_j)$ , employing the MMA
  - 17:     **end for**
  - 18:     Renumber:  $\{x_j^{(i),k}\}_{k=1, i=1}^{Np_0, 1/p_0} \mapsto x_j^{(1)}, \dots, x_j^{(N)} \sim \pi(\cdot|F_j)$
  - 19:     **for**  $i = 1, \dots, N$  **do**
  - 20:         **if**  $G^i = G(x_j^{(i)}) > b$  **then**
  - 21:              $N_F(j) \leftarrow N_F(j) + 1$
  - 22:         **end if**
  - 23:     **end for**
  - 24: **end while**
  - 25: Set  $\hat{P}_f = p_0^j \frac{N_F(j)}{N}$
  - 26: **return**  $\hat{P}_f$
- Output:**
- 27: ▶ Probability of failure obtained through the SubSim algorithm,  $\hat{P}_f$
- 

### 3.4 Practical examples of subset simulation

This section illustrates the employability of the subset simulation algorithm in geotechnical engineering through two practical examples. Firstly, a parabolic performance function of two random variables is analyzed. This example will show how the SumSim algorithm propagates samples towards the domain of failure. Secondly, a practical geotechnical example is studied. In particular,

the reliability of a square shallow foundation is analyzed. Both examples contrast the results and efficiency of both SubSim algorithm and MCS.

### 3.4.1 Parabolic performance function

Let us consider the following parabolic performance function (example adapted from [Phoon, 2008](#)):

$$g(R, S) = (R - 11)^2 - (S - 6)$$

where  $R$  denotes the resistance force, such that  $R \sim \mathcal{N}[8.5, 0.707]$ , and  $S$  denotes the external force, such that  $S \sim \mathcal{N}[5, 0.707]$ , and  $\mathcal{N}$  stands for the normal probability distribution. The system fails when the output of the performance function is equal or less than zero, i.e.  $g(R, S) \leq 0$ . It is required to compute the system probability of failure through MCS and SubSim methods.

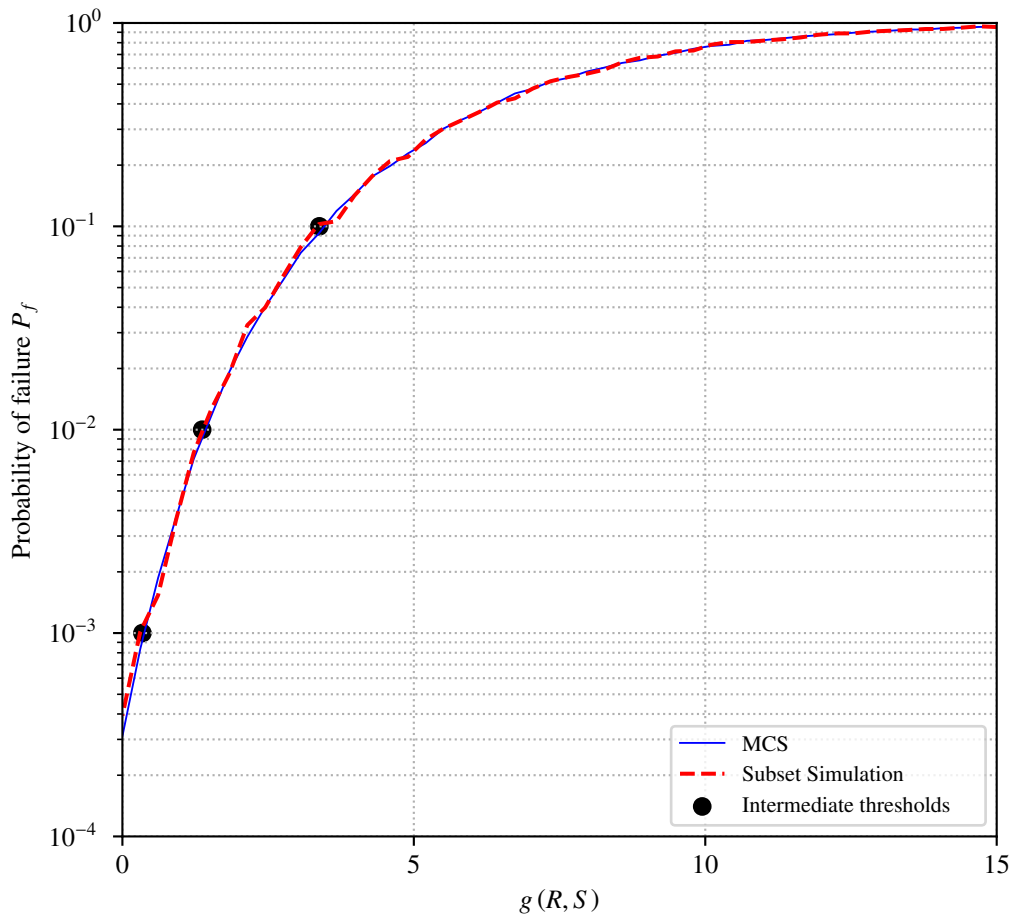
To develop this example, MCS is directly applied to the performance function with  $10^6$  samples drawn from  $R$  and  $S$ . It is expected that this number of runnings guarantees a good level of accuracy. On the other hand, the SubSim algorithm will employ 1000 samples in each simulation level. Additionally, the intermediate thresholds are defined adaptively for a conditional probability equal to  $p_0 = 0.1$ . Finally, a uniform distribution centered at  $\theta_i$  with a width equal to 2.0 is selected as the proposal distribution.

Taking into account these considerations, MCS and SubSim algorithms are executed. The probability of failure obtained by MCS was equal to  $3.38 \times 10^{-4}$ , and the respective elapsed time to compute it was equal to 170.96 seconds. On the other hand, the probability of failure obtained by the SubSim algorithm was equal to  $3.14 \times 10^{-4}$ , the total number of samples employed equal to  $N = 4 \times 10^3$ , and the respective elapsed time to compute it was equal to 14.94 seconds. It is worth noting that the SubSim algorithm defines three intermediate thresholds equal to  $b_1 = 3.39$ ,  $b_2 = 1.41$ , and  $b_3 = 0.36$ , and hence four intermediate failure events, each one with  $10^3$  samples.

From this toy model, the advantages of the SubSim algorithm over the conventional MCS are evident. Both probabilities of failure estimates are on the same order of magnitude, and for practical purposes are equal. Nonetheless, the SubSim algorithm demonstrates to be approximately 11.5 times faster than the crude MCS in this particular example.

To complement these results, figure 3-5 presents a graphical comparison of the probabilities of failure obtained using SubSim and MCS algorithms for different critical thresholds  $b \in [0, 15]$ . The similarity between the two curves is evident. There are some minor discrepancies between these two methodologies when the probability of failure is quite small. However, this behavior is expected and do not represents any drawback for practical purposes. Therefore, the accuracy of the SubSim algorithm is demonstrated.

Additionally, figure 3-6 shows how subset simulation propagates samples in the model domain. Unlike MCS, samples generated by the SubSim procedure are routed to the failure domain. There-



**Figure 3-5:** Example 3.4.1: evolution of the probabilities of failure obtained for the parabolic performance function from equation 4.2.1.3, with details of the intermediate thresholds  $b_1 = 3.39$ ,  $b_2 = 1.41$  and  $b_3 = 0.36$ .

fore, the bulk of the simulation is not wasted in the safe domain, such as in MCS, but rather efforts are focused on characterizing the failure region. Note that intermediate threshold are also plotted in figure 3-5 and figure 3-6. By defining these threshold adaptively, intermediate failure regions  $F_i$  are not equal, but rather are adjusted in a way that favors the efficiency of the algorithm.

Finally, it is worth mentioning that the problem of efficient may not be evident in a simple model with few random variables or with a simple limit state equation, such as the studied in this example. In this particular case, it may seem a trifle that Monte Carlo simulation takes approximately 2.6 minutes more than the methodology of the subset simulation. However, when the complexity of the problem increases, so the computational efforts, and hence efficiency begins to play a fundamental role. For this cases, subset simulation is a specially useful algorithm and it must be employed whenever possible.



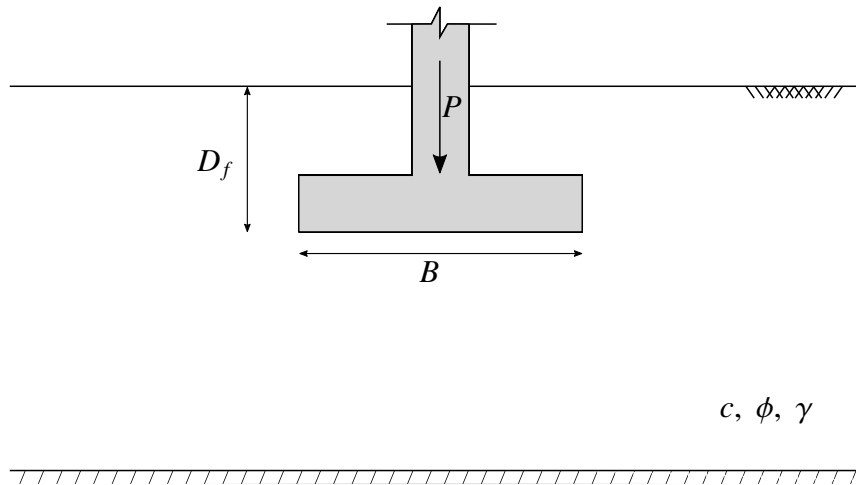
**Figure 3-6:** Failure regions and samples propagation for the SubSim algorithm in the example 3.4.1

### 3.4.2 Shallow footing with variable width

Let us consider a square shallow footing resting on a cohesive soil of stiff consistency. The foundation is located in a horizontal terrain with no groundwater table present. The material adjacent to the footing follows the Mohr-Coulomb failure criterion, i.e. it is a  $c - \phi$  soil. Finally, a unique concentric vertical load  $P$  is applied to this foundation. A scheme of this model is presented in figure 3-7.

From figure 3-7, the depth of the footing  $D_f$  is equal to 2.0 meters and the width of the footing  $B$  will be variable between 1.0 meter and 3.5 meters, i.e.  $B \in [1.0 \text{ m}, 3.5 \text{ m}]$ . The remaining parameters are considered as random variables. The applied load  $P$  follows a triangular distribution  $\mathcal{T}$  bounded on [900 kN, 1500 kN], and with a mode of 1100 kN. Soil cohesion  $c$  follows a log-normal distribution  $\mathcal{LN}$  with  $\mu_c = 10 \text{ kN/m}^2$  and  $\sigma_c = 5 \text{ kN/m}^2$ . Friction angle  $\phi$  follows a log-normal distribution  $\mathcal{LN}$  with  $\mu_\phi = 15^\circ$  and  $\sigma_\phi = 4.5^\circ$ . Soil unit weight  $\gamma$  follows a normal distribution  $\mathcal{N}$  with  $\mu_\gamma = 18 \text{ kN/m}^3$  and  $\sigma_\gamma = 1.0 \text{ kN/m}^3$ . All these random variables are considered independent.

Taking these considerations into account, it is required to estimate the probabilities of failure for the footing when varying its width  $B$ . For this purpose, MCS and SubSim algorithms must be employed, and their results contrasted.



**Figure 3-7:** Square shallow footing scheme (example 3.4.2)

To solve this problem, the equations postulated by Terzaghi (1943) will be employed to estimate the ultimate bearing capacity of the foundation  $q_u$ . Thus, the system performance function is expressed as follows:

$$g(c, \phi', \gamma) = q_u - \frac{P}{B^2}$$

Such that  $q_u$ , for square shallow footings, is equal to:

$$q_u = 1.3c'N_c + \gamma D_f N_q + 0.4\gamma B N_\gamma$$

where  $N_q$ ,  $N_c$  and  $N_\gamma$  are the Terzaghi's bearing capacity factors, defined by (Terzaghi, 1943):

$$N_q = \frac{a^2}{2 \cos^2(\pi/4 + \phi'/2)}; \quad a = \exp[(3\pi/4 - \phi'/2) \tan(\phi')]$$

$$N_c = \cot(\phi') [N_q - 1]$$

$$N_\gamma = \frac{\tan(\phi')}{2} \left[ \frac{K_p}{\cos^2(\phi')} - 1 \right]; \quad K_p = \tan^2(\pi/4 + \phi'/2)$$

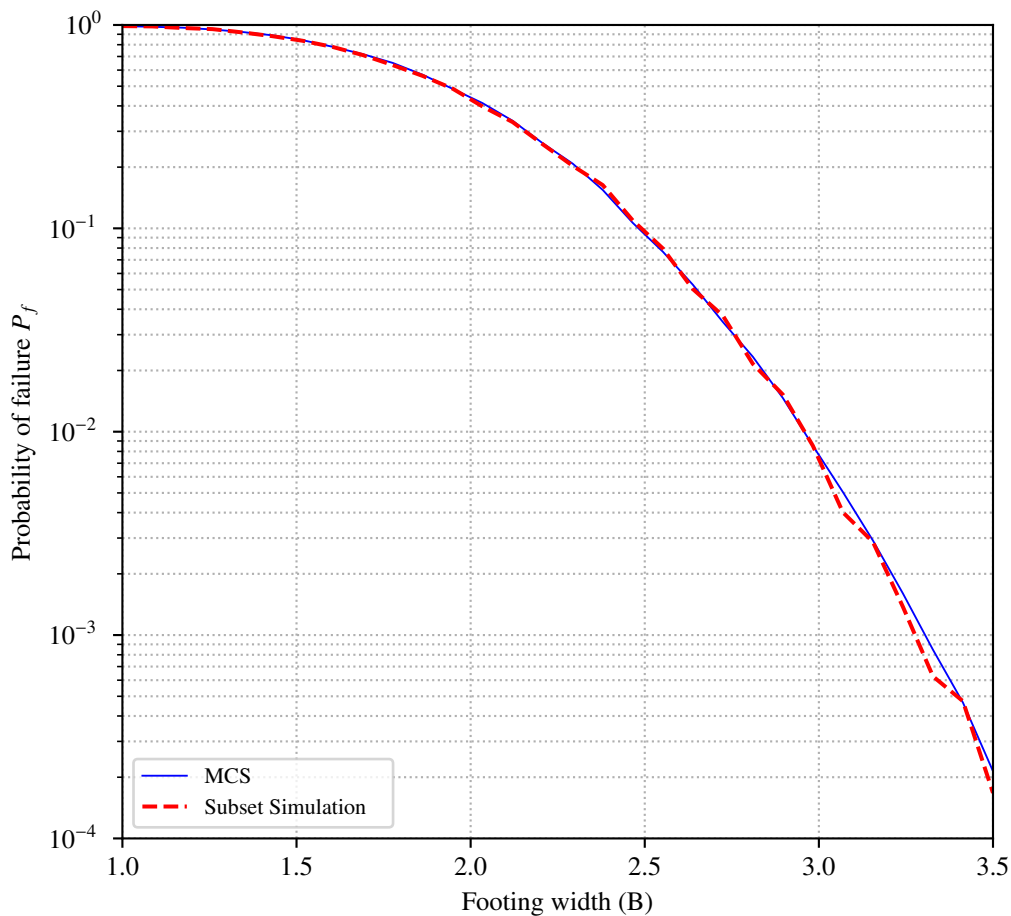
Considering this information, we can proceed to apply SubSim and MCS algorithms. In the first instance, MCS will be directly applied to the performance function using  $10^6$  samples. On the other hand, SubSim algorithm will employ 1000 samples on each simulation level, and its intermediate thresholds are defined adaptively for a conditional probability of  $p_0 = 0.1$ . Additionally, the uniform distribution centered at  $\theta_i$  with a width equal to 2 is selected as the proposal distribution for the SubSim procedure.

After executing both algorithms for values of  $B$  in  $[1, 3.5]$ , figure 3-8 is obtained. In this figure, the probabilities of failure obtained applying MCS and SubSim are represented by a blue curve and a red-dotted curve, respectively.

From figure 3-8, the similarity of the two curves is evident. Although there are some minor discrepancies when the probability of failure approaches to its smaller values, this is only a trifle in practice and do not constitute any drawback of the method. On the other hand, the efficiency of the SubSim algorithm is overwhelming. While subSet simulation required 165.26 seconds to carry out all the computations of the probability of failure for the different footing widths, MCS required 2310.78 seconds. Therefore, for this particular example SubSim was approximately 14 times faster than MCS.

For the particular width of  $B = 3.5$  m, which leads to the higher reliability, the probability of failure obtained by means of SubSim algorithm is equal to  $8.0 \times 10^{-3}$ , while the one of MCS is  $7.9 \times 10^{-3}$ . So, the discrepancy between these two values is less than 1.00%, which corroborates the SubSim accuracy. Furthermore, SubSim required only  $3 \times 10^3$  samples and 17.2 seconds for the computation of this probability of failure, while MCS required  $1 \times 10^6$  samples and 171.41 seconds. This latter fact highlight again the efficiency of the subset simulation algorithm.

It is worth mentioning that it is also possible to further improve the accuracy of the subset simulation algorithm. For this purpose, some authors take advantage of the efficiency of the algorithm and execute it  $N_{sim}$  number of times and then average the result of these simulations. The number  $N_{sim}$  is variable and it will depend on the particular conditions of each model; however, it is found in the literature that a value of  $N_{sim} = 50$  is usually adopted (Zuev et al., 2012; Alvarez et al.,



**Figure 3-8:** Probabilities of failure of example 3.4.2 computed for different footing widths (B)

2018). Finally, this procedure allows obtaining an expected value of the probability of failure and also computing some statistical parameters such as covariance of the results.

Note that in this example the simulations take longer than in the example 3.4.1. This is mainly due to the number of random variables and the complexity of the limit state equation. It is worth highlighting again that the major advantages of the SubSim algorithm appear when computing small failure probabilities or/and when the models are complex or in high-dimensions. Additionally, note that the SubSim algorithm is quite general and can be applied to any geotechnical model for estimating its reliability. This document presents the example of a simple shallow footing only for simplicity purposes, but the algorithm can be extended to more complex geotechnical systems with no problem.

# 4 On the dependence and the use of copulas in geotechnics: a state-of-the-art review

Copulas are functions that couple several marginal distribution functions through a dependence structure, in order to generate joint probability distribution functions. Copulas have been used extensively in financial applications for evaluating dependence among random variables. Recently, copula theory has been widely studied from an engineering perspective, and several applications have been found in areas like hydrology, structures or reliability. Particularly in the field of geotechnical engineering, copula theory has drawn the attention of researchers due to its advantages when modeling random variables and their dependence. In fact, several studies have been conducted in recent years evaluating the applicability of copula theory in geotechnics, and the results are conclusive about the importance and the need of its use. Therefore, this chapter is aimed to review the state-of-the-art of copula theory in geotechnical engineering, while discussing its advantages, disadvantages, limitations, applications, among other aspects, based on the existing literature.

## 4.1 Contextualization

Efforts have been made to include a probabilistic approach in the daily work of geotechnical engineering. Methodologies like load and resistance factor design (LRFD) and nominal factors of safety have been developed to integrate uncertainty into design (see, e.g., [Orr, 2000](#); [Forrest and Orr, 2010](#)). Nevertheless, complete probabilistic handling of uncertainty has not been extensively used due to two main factors: the first one is that most codes and standards are still based on deterministic factors of safety, and the second one is that a probabilistic assessment of geotechnical systems requires an exhaustive characterization of the involved variables, which is rarely achieved in practice due to high costs or lack of time.

Additionally, when a probabilistic description of the system is employed, all variables involved in the system, such as shear strength, unit weight or peak ground acceleration, among others, must be modelled as random variables. Those random variables are grouped into a vector  $\mathbf{X} =$

$[X_1, X_2, \dots, X_d]$  and represented either by a cumulative distribution function (CDF)  $F_X$  or by a probability distribution function (PDF)  $f_X$ .

For a complete probability description of uncertainty in geotechnical problems and designs, it is necessary to have a well specified joint cumulative distribution function (JCDF),  $F_X(\mathbf{X})$ , or a fully described joint probability density function (JPDF),  $f_X(\mathbf{X})$ , that relate all the random variables  $\mathbf{X} = [X_1, X_2, \dots, X_d]$  involved in the system.

Nonetheless, this level of accuracy is not generally achieved in geotechnical engineering practice, where scarce information is common. In fact, it is usual that the only probabilistic description of the involved random variables are their marginal PDFs and a correlation matrix. Hence, a joint probability distribution cannot be determined uniquely (see, e.g., [Der Kiureghian and Ditlevsen, 2009](#); [Beer et al., 2013](#)).

This lack of information and accuracy lead to assumptions that are not always well supported and that generate errors and dispersions in results, which represent a great disadvantage of probabilistic methods in practice. Some of the most usual assumptions that are chosen in order to simplify calculations are: the modelation of variables as gaussian without any evidence, the assumption that Pearson's rho coefficient can accurately model the correlation between variables, or that the variables are independent even if these are obtained from the same sample or/and the same laboratory test (e.g., shear strength parameters,  $c$  and  $\phi$ ).

Such problems widen the gap between theoretical and practical applications of probabilistic methods in the framework of geotechnical engineering. It is impossible to appropriately describe a set of geotechnical random variables when there is only one or two options of JPDFs available, e.g. multivariate normal or independence. Attempts to overcome this disadvantage have been developed, being one of the most notable the Nataf transformation ([Nataf, 1962](#)), which will be introduced in section 4.3.1.

However, most of these alternatives are quite restrictive and are based on assumptions and simplifications that affect the accuracy of the final model and that do not represent the real conditions. It was not until the emergence and popularization of copula theory that it was possible to have a comprehensive framework on dependence and construction of multivariate probability models. Copula theory managed to encompass many of the dependence methodologies that were previously employed, and it is currently consolidated as the dependence theory par excellence.

The term copula was first employed in the context of math and probability theory by [Sklar \(1959\)](#), although in reality these functions appear in earlier works that named them in a different way. A copula is a function that joins or "couple" several marginal distribution functions with the aim of constructing a multivariate probability distribution. Thus, a copula contains all the dependence information from a JPDF-JCDF, and for that reason it is also common to hear that it is called a dependence function.

In this way, by employing copula theory it is possible to construct multivariate probability dis-

tributions, where marginal distributions and dependence can be studied separately. This means a breakthrough when modeling the dependence of two or more random variables since there is a wide range of copula functions (i.e., dependence structures), and hence a wide range of multivariate models can be built. Copulas have been widely applied in areas like actuary and finance (see, e.g., McNeil et al., 2015; Cherubini et al., 2004), hydrology (see, e.g., Chen et al., 2013; Ghosh, 2010; Zhang and Singh, 2006), structural engineering (see, e.g., Goda, 2010), and more recently, in geotechnics. For instance, copulas have been used in geotechnical engineering for modeling dependence among shear strength parameters, to estimate settlement of foundations, or representing different soil properties.

Nonetheless, although copula theory has been used in different contexts and problems of geotechnical engineering, these works have been rather sparse and independent. Therefore, there is currently no methodology or framework that guides students, researchers, or consulting engineers in the area of geotechnics on the application of copulas in their particular problems. There is then the need to carry out a work that compiles the findings and conclusions of all the previous investigations on the use of copula theory in geotechnics, and that in turn serves as a basis for future practical applications and investigations of this theory in our area.

Thus, this chapter is aimed to fill this gap in the *state-of-the-art* by introducing copula theory and highlighting its potential, scope, and use in geotechnical engineering. For this purpose, it will be explained what copula functions are, when they are used, how they can be built, how they have been used in geotechnics, and what implications their use has. Obviously, this procedure involves a thorough review of the current *state-of-the-art* on the theory of copulas and its use in geotechnics.

After finalizing this chapter, questions like: do we necessarily have to model dependence in geotechnics?, what the applications of copula functions in geotechnics are there?, do copula models perform better than the traditionally Nataf transformation?, how do the gaussianity or independence assumptions perform concerning copula functions?, is there “the most suitable” copula for most geotechnical models?, can copulas be used hereinafter in geotechnical engineering practice?, what enhancements can be performed by their use?, is there any future work of the framework of copulas in geotechnical engineering?, will be answered.

The remaining of this document is organized as follows: section 4.2 introduces the mathematical concepts behind copula theory; section 4.3 exposes how dependence has been commonly addressed in geotechnics, and how traditional approaches to deal with dependence have been applied; section 4.4 explains the basis for the construction of copula functions with an emphasis in geotechnical models; in section 4.5, an *state-of-the-art* on the applicability of copula theory in geotechnical engineering is presented; section 4.6 deals with the topic of uncertainties in the copula construction and selection; in section 4.7, the applicability of copulas in reliability analysis of geotechnical models is studied and a practical example is conducted.

## 4.2 Introduction to copula theory

This section aims to introduce the basic concepts of copula theory, mainly focused on practical concepts that are needed in order to apply copula functions in the context of geotechnical engineering, after [Nelsen \(2007\)](#), [Embrechts et al. \(2001\)](#), [Embrechts et al. \(2002\)](#) and [McNeil et al. \(2015\)](#).

### 4.2.1 Preliminary concepts

This section presents some preliminary concepts that will help us later when the copula functions are formally defined.

#### 4.2.1.1 Increasing functions

A function  $f : \mathbb{R} \mapsto \mathbb{R}$  is called *increasing* if for all  $x \leq y$  the relationship  $f(x) \leq f(y)$  is satisfied. Sometimes these functions are also referred as *non-decreasing*. Furthermore, the function  $f$  is said to be *strictly increasing* if for all  $x < y$  the relationship  $f(x) < f(y)$  holds. Analogously, these function are also referred as *strictly non-decreasing* function.

On the other hand, if the increasing (or strictly increasing) function  $f$  is differentiable on the interval  $(a, b)$ , then the function  $f$  is also said to be *monotonic*.

#### 4.2.1.2 Quasi-inverses

The concept of *quasi-inverse* is a generalization of the concept of the inverse of a function, in which the inverse is defined even when the function has jumps or discontinuities. Let us consider a non-decreasing function  $F : \mathbb{R} \rightarrow R$  with  $R \subseteq \mathbb{R}$ , and the extended real line  $\bar{\mathbb{R}} = [-\infty, \infty]$ . Then, a quasi-inverse of  $F$  is any function  $F^{(-1)} : R \rightarrow \bar{\mathbb{R}}$ , such that:

- If  $t \in R$ , then  $F^{(-1)}(t)$  is any number  $x$  in  $\bar{\mathbb{R}}$  such that  $F(x) = t$ .
- If  $t \notin R$ , then  $F^{(-1)}(t) = \inf \{x \in \bar{\mathbb{R}} | F(x) \geq t\}$ .

Note that the concept of quasi-inverse and inverse of a function coincide when  $F$  is a strictly monotonically increasing function. In this case, the inverse is denoted as  $F^{-1}$ .

#### 4.2.1.3 $V_H$ -volume

Let  $S_i = [a_i, b_i]$  with  $i = 1, 2, \dots, d$ , be a non-empty subset of  $\bar{\mathbb{R}}$ , and let  $H$  be a mapping  $H : S \rightarrow \mathbb{R}$ . Furthermore, let  $B = \times_{i=1}^d S_i$  be a  $d$ -dimensional box whose vertices belongs to  $S$ . Then, the

$H$ -volume of  $B$ ,  $V_H(B)$ , is defined as the  $d$ th-order difference of  $H$  on  $B$ , as follows:

$$V_H(B) = \Delta_{a_d}^{b_d} \Delta_{a_{d-1}}^{b_{d-1}} \dots \Delta_{a_1}^{b_1} H$$

where the first order differences of  $H$  are defined as:

$$\Delta_{a_k}^{b_k} = H(t_1, \dots, t_{k-1}, b_k, t_{k+1}, \dots, t_d) - H(t_1, \dots, t_{k-1}, a_k, t_{k+1}, \dots, t_d)$$

for  $k = 1, \dots, d$ .

The above expression can be rewritten more explicitly as:

$$V_H(B) = \sum_{i_1=1}^2 \dots \sum_{i_d=1}^2 (-1)^{i_1 + \dots + i_d} H(x_{1i_1}, \dots, x_{di_d})$$

where  $x_{j1} = a_j$  and  $x_{j2} = b_j$  for all  $j = 1, \dots, d$  (see, e.g., [Embrechts et al., 2002](#)).

For instance, in a bidimensional case,  $d = 2$  and  $B = [a_1, b_1] \times [a_2, b_2]$ , and hence the  $H$ -volume of the rectangle  $B$  is the second order difference of  $H$  on  $B$ :

$$V_H(B) = H(a_2, b_2) - H(a_1, b_2) - H(a_2, b_1) + H(a_1, b_1)$$

Note that the  $d$ -dimensional function  $H$  is  $d$ -increasing if  $V_H(B) \geq 0$  for all  $d$ -boxes whose vertices fall in the domain of  $H$ .

#### 4.2.1.4 Multivariate cumulative distribution functions

A  $d$ -dimensional cumulative distribution function ( $d$ -dimensional CDF) is defined as a function  $H : \bar{\mathbb{R}} \mapsto [0, 1]$  that fulfills the following properties:

- $H$  is  $d$ -increasing.
- $H(\mathbf{t}) = 0$  if  $\mathbf{t} = [t_1, \dots, t_k, \dots, t_d] \in \bar{\mathbb{R}}$  has any element  $t_k = -\infty$ , for  $k = 1, \dots, d$ .
- $H(\infty, \dots, \infty) = 1$

#### 4.2.1.5 $k$ -margins

Let us consider the joint cumulative distribution function  $H : S_1 \times S_2 \times \dots \times S_d \rightarrow [0, 1]$ . Then, the function

$$H_j(x) = H(\sup S_1, \dots, \sup S_{j-1}, x, \sup S_{j+1}, \dots, \sup S_d)$$

is a univariate margin of  $H$ , or simply a margin of  $H$ . Note that higher dimensional margins ( $k$ -margins,  $k \geq 2$ ) can be defined by fixing less positions in the previous equation.

### 4.2.2 Copula functions

Copulas are functions that couple several CDFs through a dependence structure in order to represent a multivariate CDF. In this sense, one can characterize and define a marginal CDF for each random variable, and then model their dependence with copulas.

Mathematically, a copula  $C$  is a  $d$ -dimensional CDF whose domain is  $[0, 1]^d$  and whose range is  $[0, 1]$ , specifically  $C : [0, 1]^d \mapsto [0, 1]$ . In this regard, every marginal of a  $d$ -dimensional copula is uniform in the interval  $[0, 1]$ . Furthermore, any copula function  $C$  has the following properties:

1.  $C$  is *grounded*. That is to say, for all  $\mathbf{u} \in [0, 1]^d$ , then  $C(\mathbf{u}) = 0$  if there exists any  $u_i = 0$ , for  $i = 1, \dots, d$ .
2.  $C$  has *margins*. That is to say, for all  $\mathbf{u} \in [0, 1]^d$ , then  $C(\mathbf{u}) = u_i$  if all coordinates of  $\mathbf{u}$  are 1 except  $u_i$ .
3.  $C$  is *d-increasing*. That is to say, for all  $[a_1, \dots, a_i, \dots, a_d], [b_1, \dots, b_i, \dots, b_d] \in [0, 1]^d$ , and taking into account that  $a_i \leq b_i$  for  $i = 1, \dots, d$ , then  $V_C(\times_{i=1}^d [a_i, b_i]) \geq 0$

For instance, if the copula is bidimensional ( $d = 2$ ) then  $B = [a_1, b_1] \times [a_2, b_2]$  and hence the condition to be 2-increasing becomes:

$$C(a_2, b_2) - C(a_1, b_2) - C(a_2, b_1) + C(a_1, b_1) \geq 0$$

Furthermore, by virtue of the Lebesgue's decomposition theorem, every  $d$ -dimensional copula can be decomposed as follows (see, e.g., [Nelsen, 2007](#)):

$$C(u_1, \dots, u_d) = A_c(u_1, \dots, u_d) + S_c(u_1, \dots, u_d)$$

where

$$A_c(u_1, \dots, u_d) = \int_0^{u_1} \dots \int_0^{u_d} \frac{\partial^d}{\partial u_1 \dots \partial u_d} C(u_1, \dots, u_d) ds_1 \dots ds_d$$

and

$$S_c(u_1, \dots, u_d) = C(u_1, \dots, u_d) - A_c(u_1, \dots, u_d)$$

such that  $A_c$  is an *absolutely continuous component* and  $S_c$  is a *singular component* of the copula.

When  $S_c = 0$  on  $[0, 1]^d$ , then the copula  $C$  is said to be *absolutely continuous*, its joint PDF exists and it is given by:

$$c(u_1, \dots, u_d) = \frac{\partial^d}{\partial u_1 \dots \partial u_d} C(u_1, \dots, u_d) \quad (4-1)$$

On the other hand, when  $A_c = 0$  on  $[0, 1]^d$ , then the copula is said to be *singular* and equation (4-1) is equal to zero,  $\frac{\partial^d}{\partial u_1 \dots \partial u_d} C(u_1, \dots, u_d) = 0$  almost everywhere<sup>1</sup> in  $[0, 1]^d$  (see, e.g., Embrechts et al., 2001).

Nonetheless, when a copula has both an absolutely continuous component and a singular component, then its joint PDF cannot be easily obtained and it is sometimes undefined.

### 4.2.3 Sklar's theorem

It was previously mentioned that copulas have the function of grouping a set of marginal CDFs in order to constitute a multivariate CDF. This characteristic is given, in fact, by virtue of the *Sklar's theorem*.

Let us consider a multivariate distribution function  $F_{X_1, X_2, \dots, X_d}$  with marginal distributions  $F_{X_i}$  for  $i = 1, \dots, d$ . Then, the Sklar's theorem dictates that there exists a copula  $C$  such that:

$$F_{X_1, X_2, \dots, X_d}(\mathbf{x}) = C(F_{X_1}(x_1), F_{X_2}(x_2), \dots, F_{X_d}(x_d)) \quad (4-2)$$

for all  $\mathbf{x} := [x_1, x_2, \dots, x_d] \in \mathbb{R}^d$ . This copula is *unique* if all marginal distributions  $F_{X_i}$  are continuous; otherwise, the copula  $C$  is uniquely defined on the cartesian product of the range of each marginal distribution, that is  $\text{range}(F_1) \times \text{range}(F_2) \times \dots \times \text{range}(F_d)$ .

Analogously, by knowing a copula  $C$  and a group of one dimensional CDFs  $F_{X_i}$ , then the  $d$ -dimensional distribution function  $F_{X_1, X_2, \dots, X_d}$  can be defined by the equation (4-2). The proof of the equation (4-2) for a bidimensional case is given by Sklar (1959), while the proof is extended to a multivariate case in Sklar (1996).

Note that equation (4-2) can be rewritten in order to obtain the inverse relation, which is given by

$$C(\mathbf{u}) = F_{X_1, X_2, \dots, X_d}(F_{X_1}^{(-1)}(u_1), F_{X_2}^{(-1)}(u_2), \dots, F_{X_d}^{(-1)}(u_d))$$

for any  $\mathbf{u} := [u_1, u_2, \dots, u_d] \in [0, 1]^d$ , and where  $F_{X_i}^{(-1)}$  is the quasi-inverse of  $F_{X_i}$ , for  $i = 1, 2, \dots, d$ .

In this way, Sklar's theorem states that a continuous multivariate probability function can be expressed through its one dimensional marginal distributions and a multivariate dependence structure. This multivariate dependence structure is in fact a copula function.

Furthermore, if the copula is absolutely continuous, then the copula PDF can be defined and the  $d$ -dimensional PDF of the random variables  $X_1, X_2, \dots, X_d$  can be expressed as:

$$f_{X_1, X_2, \dots, X_d}(\mathbf{x}) = c(u_1, \dots, u_d) f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_d}(x_d) \quad (4-3)$$

where  $c(u_1, \dots, u_d)$  is given by the equation (4-1), and  $f_{X_i}$ , for  $i = 1, \dots, d$ , are the marginal distributions functions.

<sup>1</sup>It is said in a probabilistic context that a property holds *almost everywhere* when the set on which this property applies encompasses nearly all possibilities.

#### 4.2.4 The Fréchet-Hoeffding bounds

Based on the previously described properties, it can be demonstrated that any copula, regardless of its family or even its particular properties, is bounded by an upper and a lower limit, which are known as the *Fréchet-Hoeffding bounds*. In this way, any copula satisfies the following inequality:

$$W_d(\mathbf{u}) \leq C(\mathbf{u}) \leq M_d(\mathbf{u})$$

where the lower limit  $W_d$  represents *countermonotonicity*, i.e. perfect negative dependence, and the upper limit  $M_d$  represents *comonotonicity*, i.e. perfect positive dependence.

In a bivariate case,  $W(u_1, u_2)$  and  $M(u_1, u_2)$  are expressed as (see, e.g., [Nelsen, 2007](#), p. 11):

$$W(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$$

$$M(u_1, u_2) = \min(u_1, u_2)$$

and in the  $d$ -dimensional case they are given by:

$$W_d(\mathbf{u}) = \max(u_1 + u_2 + \dots + u_d - d + 1, 0)$$

$$M_d(\mathbf{u}) = \min(u_1, u_2, \dots, u_d)$$

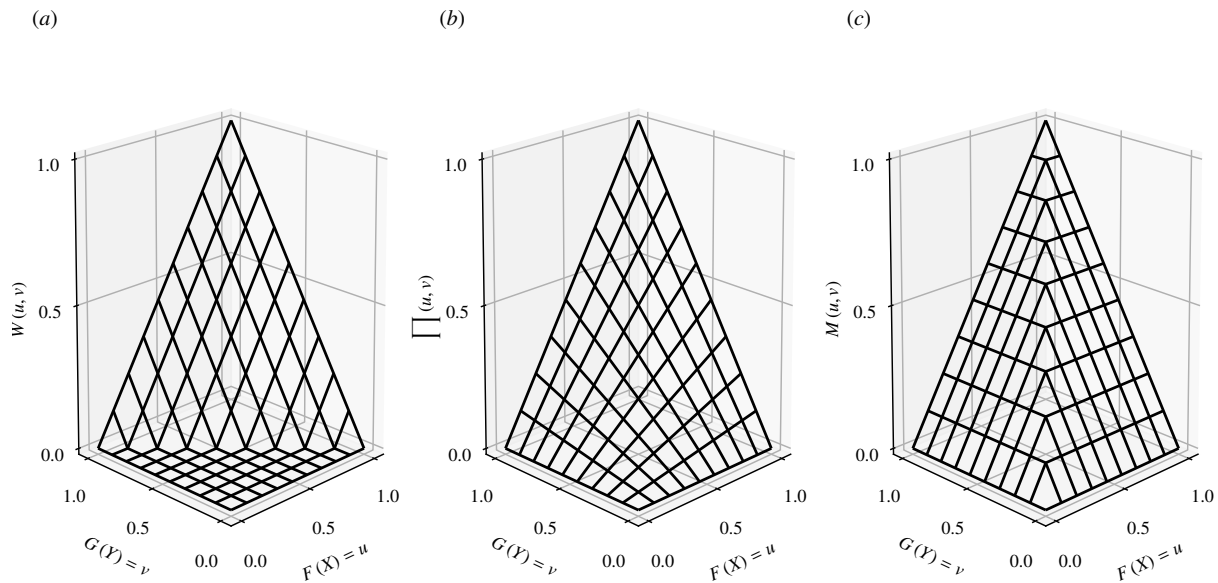
Another structure of dependence worth mentioning is in fact the one that represents independence. From probability theory, it is well known that the probability of occurrence of independent events is equal to the product of the probability of each particular event. Thus, by extending this concept to copula theory, independence is represented as:

$$\Pi_d(\mathbf{u}) = \prod_{i=1}^d u_i = u_1 u_2 \cdots u_d$$

Note that both  $M_d$  and  $\Pi_d$  are copulas for two or higher dimensions, and these are known as the comonotonic and the product/independence copula, respectively. On the other hand,  $W_d$  can only be considered a copula in a two-dimensional case, and it is known as the countermonotonic copula  $W$ , since for higher dimensions it does not fulfill the requirements to be a copula function. Nonetheless, the Fréchet-Hoeffding lower bound still holds for three and more dimensions, even if the countermonotonic copula  $W$  does not exist (see [Nelsen, 2007](#), p. 47).

For a better understanding of these structures, figure 4-1 presents the *Fréchet-Hoeffding bounds* in a two-dimensional case, as well as the independence copula.

In this sense, any copula function must lie inside the polyhedron formed by the Fréchet-Hoeffding bounds. There are some copulas capable of approaching both countermonotonicity and comonotonicity, thus covering a wide spectrum of dependence values. These copulas are known as *comprehensive copulas* and are especially useful to model random variables whose dependence is



**Figure 4-1:** Fréchet-Hoeffding bounds and independent copula in two dimensions. (a) counter-monotonic copula  $W$ , (b) independence copula  $\Pi$ , (c) comonotonic copula  $M$ .

strongly positive or negative. Some examples of comprehensive copulas are the Gaussian Copula, the Plackett copula, or the Frank copulas.

Finally, although there are certain restrictions for a function on  $[0, 1]^d$  to be considered as a copula function, there are actually many options of copulas reported in the literature. Each copula has its particular properties and is suitable for different conditions. Furthermore, some copulas share certain properties or are derived in a similar way, so it is common to group them in some families or classes. The fact that there are many copulas in literature is a great advantage in comparison to the independence or gaussianity assumption. With copulas it is possible to model different dependence structures where, as a plus, marginal distribution functions can be selected independently.

## 4.2.5 Some copula properties

In this section, some properties of the copula functions are discussed, in particular the range of dependence values, radial symmetry and tail dependence. These copula properties are useful when evaluating the suitability of a copula for modeling different dependence structures between random variables.

### 4.2.5.1 Range of dependence values

As stated before, any copula function is bounded by the Fréchet-Hoeffding bounds. However, not all copulas have the ability to approach both limits, but are restricted to a narrower range. In other

words, every copula is capable of modeling a certain spectrum of dependence, which will never be outside the Fréchet-Hoeffding bounds but will be a subset of them.

Copula functions with the property of approaching both limits are known as *comprehensive copulas*. Some examples of copulas with this property are the Gaussian, Student- $t$ , Plackett, or Frank copulas. On the other hand, some other copulas can only approach one of the two limits; for instance, Gumbel or Joe copula can approach the comonotonic copula  $M_d$ , meanwhile the bivariate Nelsen No. 16 copula can approach the countercomonotonic copula  $W$ . Finally, some other copulas are not capable approaching any of the Fréchet-Hoeffding bounds, such as the Ali-Mikhail-Haq copula, which is restricted to a narrow range of weak dependence structures (Nelsen, 2007; McNeil et al., 2015)

In this way, it is important to take this property into account when evaluating the suitability of a copula to model certain dependence structures. For example, if the dependence between random variables is known to be strongly negative, the Gumbel, Joe, or Ali-Mikhail-Haq copulas would not be a good choice for modeling this dependence. More details of these characteristics for certain copulas of interest will be presented in section 4.4.

#### 4.2.5.2 Radial symmetry

It is said that a random vector  $\mathbf{X} \in \mathbb{R}^d$  (or its distribution function) is radially symmetric about  $\mathbf{a} \in \mathbb{R}^d$  if:

$$\mathbf{X} - \mathbf{a} \stackrel{d}{=} \mathbf{a} - \mathbf{X}$$

where  $\stackrel{d}{=}$  stands for the *equality in distribution* property<sup>2</sup>.

Analogously, a copula is said to be radially symmetric if the random vector  $\mathbf{U} \in [0, 1]^d$ , where  $\mathbf{U}$  has a distribution function  $C$ , has the following property (McNeil et al., 2015, chapter 5):

$$\mathbf{U} \stackrel{d}{=} \mathbf{1} - \mathbf{U}$$

Examples of radially symmetric copulas are Gaussian and Student- $t$  copulas, while examples of non-radially symmetric copulas are Gumbel and Clayton copulas.

There are other concepts of symmetry in copula theory, such as the *marginal symmetry* or the *joint symmetry*; however, these properties are not exposed in this document, so readers interested in these concepts are referred to Nelsen (2007).

---

<sup>2</sup>If the sample space is a subset of the real line  $\mathbb{R}$ , then two random variables  $X$  and  $Y$  are said to be equal in distribution  $X \stackrel{d}{=} Y$  when they have the same distribution function, that is  $P(X \leq x) = P(Y \leq x)$  for all  $x$ .

### 4.2.5.3 Tail dependence

Let us consider two random variables,  $X$  and  $Y$ , with marginal CDFs  $F_X$  and  $F_Y$ , respectively; these two random variables are associated through a copula  $C$ . Coefficient of upper tail dependence  $\lambda_U$  and coefficient of lower tail dependence  $\lambda_L$ , are defined as follows (Nelsen, 2007):

$$\lambda_U = \lim_{q \rightarrow 1^-} P(F_Y(y) > q | F_X(x) > q) \quad (4-4)$$

$$\lambda_L = \lim_{q \rightarrow 0^+} P(F_Y(y) \leq q | F_X(x) \leq q) \quad (4-5)$$

Conceptually, *upper tail dependence coefficient*  $\lambda_U$  is the conditional probability that  $Y$  exceeds the quantile  $q$  given that  $X$  also exceeds the same quantile  $q$ , as the quantile approaches to 1. Analogously, *lower tail dependence coefficient*  $\lambda_L$  is the conditional probability that  $Y$  be inferior to the quantile  $q$  given that  $X$  also is smaller then the same quantile  $q$ , as the quantile approaches to 0.

Assuming that both  $X$  and  $Y$  are continuous, equations (4-4) and (4-5) can be expressed as (Nelsen, 2007):

$$\lambda_U = \lim_{q \rightarrow 1^-} \frac{1 - 2q + C(q, q)}{(1 - q)} \quad (4-6)$$

$$\lambda_L = \lim_{q \rightarrow 0^+} \frac{C(q, q)}{q} \quad (4-7)$$

Equations (4-6) and (4-7) show that the coefficients of tail dependence are sole properties of the copula, and not of the marginal distributions. Furthermore, the existence or absence of one of these coefficients does not imply the existence or absence of the other. In other words, a copula  $C$  may have both upper and lower tail dependence, only one of those coefficients or none.

It is important to evaluate the dependence structure in the tails of the distributions since the failure domain is usually located there, and hence tails have a great impact on the probability of failure (Ditlevsen and Madsen, 1996).

For this reason, coefficients of tail dependence are key measures of dependence in copula theory. Coefficients of tail dependence aim to quantify the associativity between random variables in the upper-right or lower-left quadrant of the  $d$ -dimensional box; or in other words, when all random variables take extreme values simultaneously.

## 4.2.6 Empirical copulas

In practice, and specifically in geotechnical engineering practice, it is common to have a small sample available from a population that is considerably much larger in the number of elements. Under these circumstances, the underlying copula (or dependence structure) of the variables is unknown, and the usual practice is to try to infer the true function from the available information.

Thus, as the amount of data in the sample increases and approaches the size of the population, the inferred copula also converges to the real copula of the analyzed multivariate data.

However, instead of inferring a copula from a reduced sample, it is also possible to construct an *empirical copula* from these multivariate data, as follows:

Let us consider a set of  $N$  observations  $\{(x_1^i, x_2^i, \dots, x_d^i) : i = 1, \dots, N\}$  from a  $d$ -dimensional random vector  $\mathbf{X} = (X_1, X_2, \dots, X_d)$ . The pseudo-copula samples can be defined as follows:

$$\hat{U}_k^i = \frac{R_k^i}{N} \quad (4-8)$$

where  $R_k^i$  is the rank of the observation  $x_k^i$ :

$$R_k^i = \sum_{j=1}^n \mathbb{I}(x_k^j \leq x_k^i) \quad (4-9)$$

Hence, the corresponding empirical copula is then defined as

$$C^N(u_1, \dots, u_d) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{U}_1^i \leq u_1, \dots, \hat{U}_d^i \leq u_d) \quad (4-10)$$

In this sense, the empirical copula can be understood as the empirical distribution of the rank transformed data. In the same way, as the sample size increases the empirical copula will approach to the true underlying copula of the multivariate data.

### 4.2.7 Vine copulas

Constructing copulas in high dimensions is not a trivial problem. Although there is a large number of copula functions in the literature, most of them are bivariate or applicable in low-dimensions. In practice, the group of copula functions in high dimensions is rather small and their properties are limited. There are mainly two drawbacks when trying to implement copulas in high dimensions:

1. There is no analytical solution for the extension to high-dimensions of the copula function, or their extension may become unfeasible.
2. The dependence structure becomes too restrictive and it is not possible to evaluate different dependence structures between pairs of variables.

For example, if one wants to build a geotechnical model with 5 random variables, there would have few copula options to choose from and one would have to assign the same dependence structure to all variables, which may be different from reality.

Fortunately, in recent years this problem has been circumvented by the so called *vine copula theory* (Bedford and Cooke, 2002). In this section, we briefly introduce it mostly based on Czado (2019)

and [Aas et al. \(2009\)](#). Nonetheless, it is worth clarifying that the explanation given here is only introductory, and an in-depth presentation can be found in [Joe \(1996\)](#), [Bedford and Cooke \(2001\)](#) and [Bedford and Cooke \(2002\)](#).

Vine copula theory states that it is possible to express a multivariate probability density function as a product of pair-copula densities and marginal densities. Vine copula theory employs bivariate copulas as building blocks for elaborating higher-dimensional distributions, where the dependence is determined by these bivariate copulas in a nested set of trees. In this way, vine copulas allows to construct models in high dimensions, in which it is possible to evaluate different structures of dependence between each pair of variables.

To clarify the above concepts, let us consider a vector of  $d$  random variables  $\mathbf{X} = [X_1, \dots, X_d]$  with a joint density function  $f(x_1, \dots, x_d)$ . It is possible to decompose this joint density function as follows:

$$f(x_1, \dots, x_d) = f(x_d) f(x_{d-1}|x_d) f(x_{d-2}|x_{d-1}, x_d) \cdots f(x_1|x_2, \dots, x_d); \quad (4-11)$$

this decomposition is unique up to a relabelling of the variables ([Aas et al., 2009](#)).

Now, observe that equation (4-3) could be expressed for a bivariate case as:

$$f(x_1, x_2) = c_{12}(F_1(x_1), F_2(x_2)) f(x_1) f(x_2)$$

and hence the conditional densities as:

$$f(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1) \quad (4-12)$$

The copula conditional density of equation (4-12) can be employed in equation (4-11) replacing each conditional term. In this way, a joint PDF can be decomposed into pair-copulas times conditional marginal densities. For instance, one possible decomposition for a three-dimensional PDF would be:

$$f(x_1, x_2, x_3) = f_1(x_1) f_{2|1}(x_2|x_1) f_{3|12}(x_3|x_1, x_2)$$

where

$$\begin{aligned} f_{2|1}(x_2|x_1) &= c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) \\ f_{3|12}(x_3|x_1, x_2) &= c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) f_{3|2}(x_3|x_2) \\ f_{3|2}(x_3|x_2) &= c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \end{aligned}$$

and hence:

$$\begin{aligned} f(x_1, x_2, x_3) &= f_1(x_1) f_2(x_2) f_3(x_3) \\ &\quad \times c_{12}(F_1(x_1), F_2(x_2)) c_{23}(F_2(x_2), F_3(x_3)) \\ &\quad \quad \quad \times c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \end{aligned}$$

Nonetheless, as stated before this decomposition is not unique, and as the number of variables in the model increases, the number of possibilities of pair-copula constructions also increases. For instance, a model of three variables has 6 possibilities of decomposition, a model of four variables has 24 possibilities of decomposition, and a model of five variables has 240 possibilities of decomposition (Aas et al., 2009).

For the sake of organizing all these options of decomposition, Bedford and Cooke (2001) introduced a graphical tool denominated *the regular vine*. A regular vine is a graphical aid for labeling bidimensional constraints in high-dimensional probability functions. When regular vines are combined with bivariate copulas, vine copulas are established.

A regular vine of  $d$  variables can be seen as a nested set of connected trees, in which the edges of the tree  $T_j$  are the nodes of the tree  $T_{j+1}$ , and the edges of the tree  $T_{j+1}$  are the nodes of the tree  $T_{j+2}$ , and so on. Furthermore, two edges in the tree  $T_j$  are joined in the tree  $T_{j+1}$  if and only if the edges share a common node. In a regular vine copula, the nodes in the first tree are the random variables, and from then on each edge and node is a conditional copula density.

Regular vines is still a quite general category that embraces several classes of pair-copula decompositions. Among all the variants of the regular vines, it is worth mentioning two special cases: the *D-vine* and the *canonical vine*. Each one of these two classes of regular vines has a specific way of describing the decomposition of a density function.

In a D-vine, each tree is a path. Therefore, a D-vine of  $d$  variables would have  $T_j$  trees, with  $j = 1, \dots, d - 1$ , and each tree would have  $d + 1 - j$  nodes and  $d - j$  edges. Note that each edge corresponds to a pair-copula density, and each edge label corresponds to the subscript of the pair copula density. Figure 4-2 presents an example of a D-vine with 5 variables.

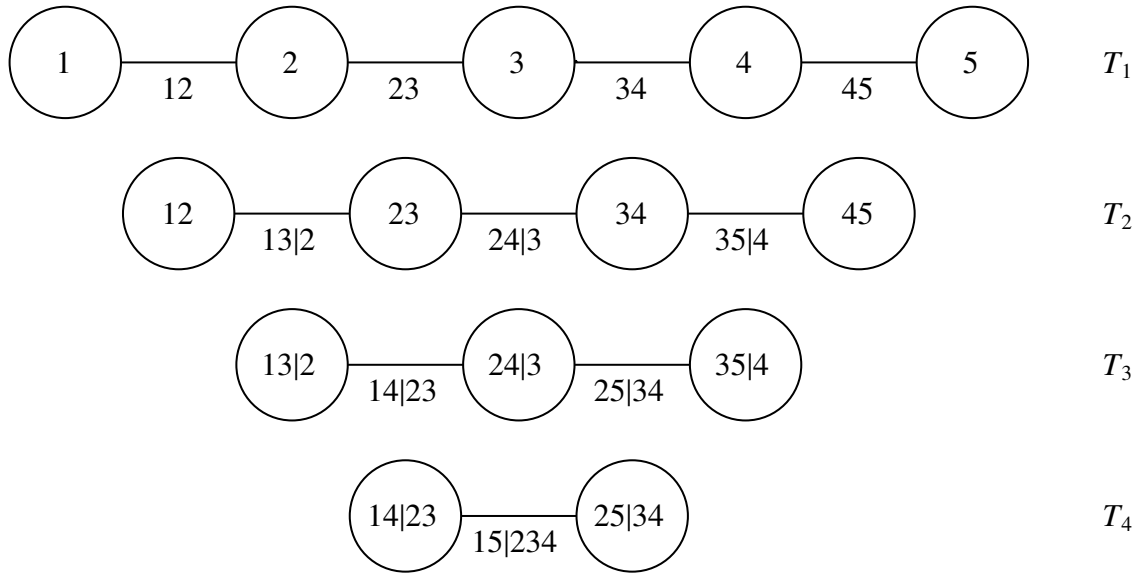
From figure 4-2 the decomposition of the 5-dimensional probability model can be expressed as:

$$f_{12345} = f_1 f_2 f_3 f_4 f_5 c_{12} c_{23} c_{34} c_{45} c_{13|2} c_{24|3} c_{35|4} c_{14|23} c_{25|34} c_{15|234}$$

The general equation in which a  $d$ -dimensional probability density function is expressed through a D-vine is given by:

$$f(x_1, \dots, x_d) = \prod_{k=1}^d f(x_k) \times \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1, \dots, i+j-1} \left( F(x_i | x_{i+1}, \dots, x_{i+j-1}), F(x_{i+j} | x_{i+1}, \dots, x_{i+j-1}) \right)$$

In a canonical vine, each tree  $T_j$  has an unique central node that connects to the other nodes through  $d - j$  edges. This type of structure is advantageous when one knows that a particular variable is the key variable that governs the interactions in the dataset. For instance, it is typical in geotechnical



**Figure 4-2:** Example of a D-vine with 5 variables, 4 trees and 10 edges. Each edge (represented as a line) is associated with a pair-copula conditional density.

engineering to relate the  $(N_1)_{60}$  of the SPT to multiple parameters and variables of soils; then, it would be desirable to recreate a multivariate model where  $(N_1)_{60}$  was at the root of the canonical vine.

Figure 4-3 presents an example of a canonical vine with 5 variables.

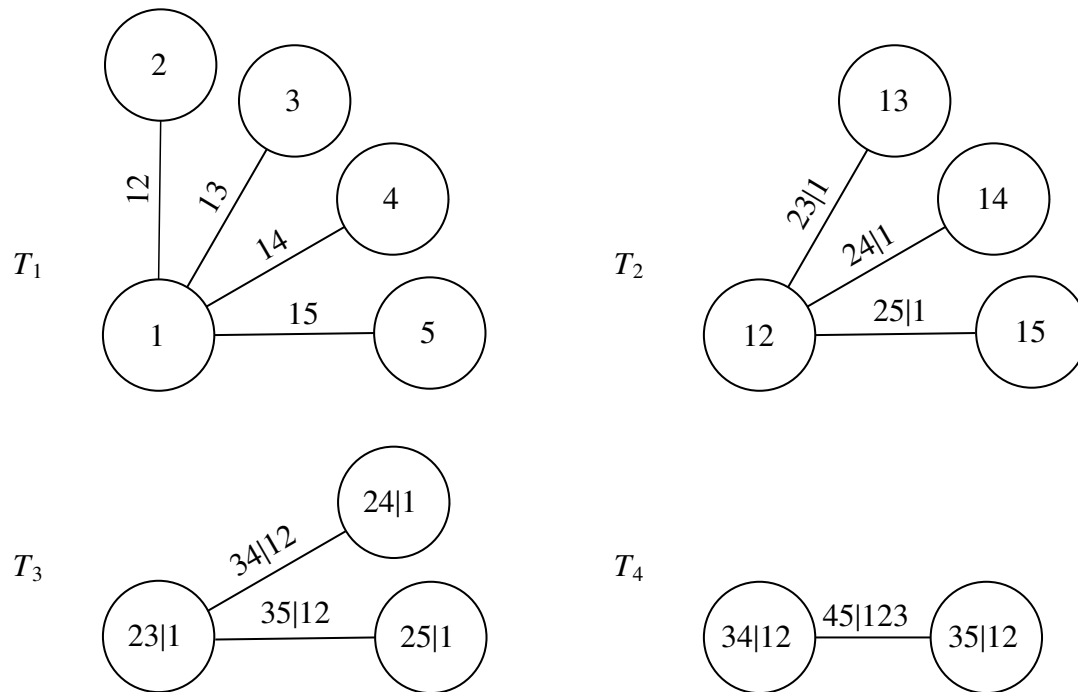
From figure 4-3 the decomposition of the 5-dimensional probability model can be expressed as:

$$f_{12345} = f_1 f_2 f_3 f_4 f_5 c_{12} c_{13} c_{14} c_{15} c_{23|1} c_{24|1} c_{25|1} c_{34|12} c_{35|12} c_{45|123}$$

The general equation in which a  $d$ -dimensional probability density function is expressed through a canonical vine is given by:

$$f(x_1, \dots, x_d) = \prod_{k=1}^d f(x_k) \times \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{j,j+i|1, \dots, j-1} \left( F(x_j | x_1, \dots, x_{j-1}), F(x_{j+i} | x_1, \dots, x_{j-1}) \right)$$

Constructing vine copulas from a dataset encompass three general steps. First, the structure of the vine must be specified. That is to say, the type of the vine and the interaction of the nodes (and variables) on each tree. Second, a bivariate copula function must be established for each edge on each tree. A wide range of bivariate parametric copula functions exist in the literature and can be employed. Third, each copula must be adjusted to the available data. Several methodologies exist in the literature for this last purpose.



**Figure 4-3:** Example of a canonical vine with 5 variables, 4 trees and 10 edges. Each edge is associated with a pair-copula conditional density.

Some of these concepts will be discussed in more detail in this document. In particular, section 4.4.3 presents several families of copulas and copula functions that are popular in geotechnics, section 4.4.4 explains different methodologies for adjusting copulas to data, and the issue of defining the best copula for a dataset is treated in section 4.4.5.

Nonetheless, it is recalled that the exposition of vine copulas given here is only introductory. More details of this theory must be consulted in the specialized references. The applicability of vine copulas in geotechnics has a great potential that is just beginning to be explored, so more detailed investigations must be carried out.

### 4.3 Some remarks about dependence in geotechnics

In order to understand the essence of copulas and their applicability in geotechnical engineering, it is necessary to clarify what dependence and correlation are, and how they are related to multivariate probability distribution functions. Although these two concepts are sometimes assumed as synonyms, from a statistical perspective they have different meanings.

On the one hand, dependence can be better approached from the concept of statistical indepen-

dence: two random variables  $X$  and  $Y$  are said to be *statistically independent* when the occurrence (or non-occurrence) of an event associated with  $X$  does not affect the occurrence (or non-occurrence) of an event associated with  $Y$ . In other words, two random variables are statistically independent if the occurrence of one event  $X$  does not have implications on the probability distribution of another event  $Y$ . This can be mathematically written through conditional probabilities as:

$$P(X|Y) = P(X) \quad \text{or} \quad P(Y|X) = P(Y)$$

thus, two random variables are considered dependent if the conditions to be statistically independent are not satisfied.

On the other hand, correlation is a statistical measure of association between two random variables. It must bear in mind that the term correlation is sometimes understood as a measure of linearity between two random variables, such as the Pearson's rho correlation coefficient does (see, e.g., [Ang and Tang, 2007](#)), although this term is also applicable to other non-linear measures of dependence, such as the rank measures that we will study in section [4.4.2](#).

In this way, dependence is a more general concept than correlation, and in fact, it encompasses all types of statistical correlations. Thus, correlation implies dependence, but dependence does not necessarily implies correlation, e.g. a Pearson's rho correlation coefficient equal to zero does not mean statistical independence between two random variables, since there could be other structures of dependence different from linearity.

Following these concepts, there are many variables in geotechnics that can be considered dependent, such as the shear strength parameters of the Mohr-Coulomb failure criterion, Atterberg limits, parameters of the soil-water characteristic curves, foundation settlement parameters, among others. Geotechnical literature is full of correlations and equations to predict values of some parameters based on some others that are easier to obtain. For instance, the  $(N_1)_{60}$  of SPT is commonly related to many other parameters such as the unconfined compressive strength  $q_u$ , cone tip resistance  $q_c$ , soil friction angle etc. Readers interested in some of these correlations are referred, for instance, to [Ameratunga et al. \(2016\)](#).

Additionally, it is found that the most used measure of correlation in geotechnical engineering corresponds to the Pearson's rho coefficient; this is expected given its popularity, simplicity, and that it is the canonical measure of dependence of the widely used multivariate normal distribution. This coefficient and its implications will be further studied in section [4.4.2](#).

Now, starting from the concept of dependence, it is necessary to construct multivariate probability distribution functions that indicate the probability of occurrence of two or more dependent events. These functions are essential to simulate random values with certain dependence structures or to carry out reliability analysis and compute probabilities of failure of geotechnical models. However, in practice there are few multivariate distribution functions that are easy to construct and extend to high dimensions, which narrows the range of possibilities when representing dependence through

a probability distribution.

Furthermore, it is unfortunately usual in geotechnical engineering practice to have a reduced amount of data due to quite limited budgets or lack of time. With this scarce information at hand, it is only possible to define marginal distributions, i.e. univariate distribution functions for each random variable, and a correlation matrix, which is commonly constructed through Pearson's rho coefficients among random variables. This condition is known as incomplete probability information, and a basic introduction to its characteristics could be found in [Der Kiureghian and Liu \(1986\)](#).

Under these circumstances, the multivariate normal distribution is commonly adopted, without validation, for representing probabilities of the joint occurrence of the dependent random variables. It is well known that this distribution has many advantages since it has been widely studied, its extension to high dimensions is straightforward, it only requires the definition of a covariance matrix based on Pearson's rho coefficients, and in general, its application is quite simple. The probability density function of a  $d$ -dimensional normal distribution is given by:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]}{\sqrt{(2\pi)^d \det \boldsymbol{\Sigma}}}$$

where  $\mathbf{x}$  is a random vector in  $\mathbb{R}^d$ ,  $\boldsymbol{\mu} \in \mathbb{R}^d$  is the mean vector, and  $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$  is the covariance matrix constructed through Pearson's rho correlation coefficients.

However, equation 4.3 is too restrictive since all random variables are assumed to be normally distributed, and dependence is limited uniquely to a Gaussian type. This results in drawbacks that appear when modeling joint occurrences of some geotechnical random variables, such as shear strength parameters, where physically both cohesion and friction angle cannot be negative, and this restriction can not be modeled using the multivariate Gaussian distribution of equation 4.3.

In consequence, there is a need of exploring other alternatives in this regard. Some early works, such as [Der Kiureghian and Liu \(1986\)](#), studied the issue of the representation of random vectors and the computations of the probabilities of failure under incomplete probability information. However, although there are several possibilities for this purpose, the truth is that very few are applicable in practice since the majority are restricted only to a bivariate case and/or only are capable of representing small correlations between random variables. One of the most widely used alternatives, until recent years when copula theory was developed, was the Nataf transformation. This transformation we will be introduced next.

### 4.3.1 The Nataf transformation

The Nataf transformation ([Nataf, 1962](#)) is an approach that appears as a solution to the problem of modeling dependence when probability information is reduced to marginal distributions of all

random variables and linear correlation coefficients among them all, which is the usual case in engineering practice, and particularly in geotechnical engineering.

Following [Liu and Der Kiureghian \(1986\)](#) and [Li et al. \(2008\)](#), the Nataf transformation can be defined as a transformation of random variables from the original/physical space to a mutually independent standard space. In this sense, a  $d$ -dimensional vector of dependent random variables  $\mathbf{X}$ , with marginal distributions  $F_{X_i}(x_i)$ , is transformed into a vector of independent random variables  $\mathbf{Z}$  that follows a standard normal probability function,  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , where  $\mathbf{I}$  stands for the  $d$ -dimensional identity matrix, as follows:  $\Phi(\mathbf{Z}) = F_{\mathbf{X}}(\mathbf{x})$ , such that

$$Z_i = \Phi^{-1}[F_{X_i}(x_i)] \quad i = 1, \dots, d \quad (4-13)$$

where  $F_{X_i}$  are the continuous marginal CDFs, and  $\Phi$  is the standard normal CDF.

Additionally, based on the Nataf transformation, the joint PDF of random vector  $\mathbf{X}$  in the original/physical space can be expressed as:

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n) \frac{\phi_n(\mathbf{z}, \boldsymbol{\rho}_0)}{\phi(z_1) \phi(z_2) \dots \phi(z_n)} \quad (4-14)$$

where

$$\phi_n(\mathbf{z}, \boldsymbol{\rho}_0) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\rho}_0)}} \exp\left(-\frac{1}{2} \mathbf{z}^T \boldsymbol{\rho}_0 \mathbf{z}\right)$$

is the  $d$ -dimensional standard normal PDF with correlation matrix  $\boldsymbol{\rho}_0$ . In general, equation(4-14) is known as *the Nataf distribution*.

It is worth noting that the correlation matrix  $\boldsymbol{\rho}_0$  is different from correlation matrix  $\boldsymbol{\rho}$ , since the former corresponds to the standard normal random vector  $\mathbf{Z}$  of equation(4-13), and the latter corresponds to the original vector  $\mathbf{X}$ . The correlation matrix  $\boldsymbol{\rho}$  will not remain constant after a non-linear transformation of the original random variables (more of this behavior is explored in section 4.4.2). However,  $\boldsymbol{\rho}_0$  can be obtained from  $\boldsymbol{\rho}$  by applying the following equation, which is presented for a two dimensional case but it can be extended to more dimensions ([Der Kiureghian and Liu, 1986](#)):

$$\rho_{ij} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( \frac{F_i^{-1}(\Phi(z_i)) - \mu_i}{\sigma_i} \right) \left( \frac{F_j^{-1}(\Phi(z_j)) - \mu_j}{\sigma_j} \right) \phi_{Z_i, Z_j}(z_i, z_j, \rho_{0,i,j}) dz_i dz_j \quad (4-15)$$

Equation(4-15) can be solved iteratively through numerical methods. Nonetheless, some semi-empirical formulas for approximating  $\boldsymbol{\rho}_0$  from  $\boldsymbol{\rho}$ , under different conditions, were presented by [Der Kiureghian and Liu \(1986\)](#). Additionally, a more general Gauss-Hermite-Quadrature method was presented by [Li et al. \(2008\)](#), which leads to better approximations of  $\boldsymbol{\rho}_0$  than the semi-empirical formulas of [Der Kiureghian and Liu \(1986\)](#).

As we can observe, with the use of the Nataf transformation we can couple dependent random variables with different univariate probability distributions, which is more flexible than the multivariate normal distribution of equation 4.3. Additionally, Nataf transformation integrates very well with popular reliability methods, such as the First Order Reliability Method (FORM) or the Monte Carlo simulation (MCS) (see, e.g., [Li et al., 2008](#)). Due to this, the Nataf transformation has been the preferred method for many years to handle dependent random variables with different univariate distributions.

However, Nataf transformation has several implicit hypotheses, and the most notable is the one of the dependence structure. In recent years, copula theory has given new insights into dependence, and it has allowed to elucidate the hidden dependence hypothesis behind Nataf transformation. In this regard, it was demonstrated by [Lebrun and Dutfoy \(2009b\)](#), in the light of copula theory, that Nataf transformation is nothing but a particular case of the Gaussian copula.

Now, what are the implications of Nataf transformation being a particular case of Gaussian copula? In fact, in the light of copula theory, Gaussian copula is just one structure of dependence that belongs to the wide range of dependence structures than can be modeled through copulas. In this way, Gaussian copula, or Nataf transformation, is one of many options and perhaps it may not be the best structure of dependence for representing a dataset, leading thus to misleading models. This topic is further explored in subsequent chapters of this document.

It is worth mentioning that, in addition to the work of [Lebrun and Dutfoy \(2009b\)](#), a generalization of the Nataf transformation to all kind of elliptical copulas was exposed by [Lebrun and Dutfoy \(2009a\)](#). In this sense, Nataf transformation was extended to represent any elliptical copula, not just a Gaussian copula. Besides this, the popular FORM/SORM were also extended to this generalized Nataf transformation.

The advantage of copulas is evident since, without them, there is only a narrow range of multivariate distribution functions in practice and those are too restrictive in terms of dependence. Additionally, copula theory is more general than the widely used Nataf transformation, and hence its range of possibilities is larger. Its multiple advantages have led copulas to be applied in many fields, being geotechnics one of them.

## **4.4 Construction of copulas for modeling dependence among geotechnical variables**

As indicated above, there is a great variety of copula functions in the literature, each with its own particular properties and special characteristics that make them suitable for different dependence conditions. However, considering this wide range of possibilities, the question arises as to which is the best choice of copula to model our variables of interest, and in particular, their dependence.

Therefore, this section will be dedicated to elucidate, from a practical perspective, how this question can be solved in the best way possible based on the available information. In other words, the concepts of the construction of copulas to model geotechnical parameters will be introduced in this section, based on the latest postulates of the state-of-the-art. Nonetheless, it is worth mentioning that the methodology presented here is not a straitjacket and different variants can be applied depending on the particular conditions of the problem of interest. In this way, the content of this section is a guide that readers can adopt as a basis for their particular problems.

In this way, five general steps are identified in the construction and definition of a copula function for the purposes of modeling geotechnical variables. They are listed in the followings:

1. Define and fit the marginal PDF of each random variable involved in the model.
2. Estimate the strength and direction of the existing dependence among the random variables.
3. Define a set of copula functions  $C$  that could fit the dependence characteristics of the random variables of interest.
4. Fit the candidate copula functions in  $C$  to the available information.
5. Select the most suitable copula function from the set of candidate copulas  $C$  through some goodness-of-fit test.

Each of these steps will be covered in more detail later. In particular, section 4.4.1 provides some references on the adjustment of random variables to one-dimensional PDFs. Section 4.4.2 introduces the concept of correlation, and explains its role as a first approximation to the magnitude and direction of dependence. Section 4.4.3 introduces some of the copula functions that, after reviewing the state-of-the-art, were identified as the most popular to model geotechnical parameters. Section 4.4.4 explains how to fit copula functions to the available data or dependence information. Finally, section 4.4.5 introduces some practical techniques for the definition of the most suitable copula among a set of candidate copulas  $C$  that seek to model the dependence of the random variables studied.

#### 4.4.1 Selection of marginal distributions

There is a wide variety of univariate distribution functions in the literature, both continuous and discrete. All of these functions can be used as marginal distributions, but in particular, we will be interested in continuous probability functions. Exposing all of these distributions, or a large number of them, is outside the scope of this document, so readers are referred to specialized references such as [Montgomery and Runger \(2010\)](#), [Ang and Tang \(2007\)](#) or [Kottegoda and Rosso \(2008\)](#). However, it is noted that some of these continuous distributions have been quite popular in geotechnics, such as the normal, lognormal and uniform probability functions. Some other continuous distributions have also shown a great potential for modeling geotechnical parameters,

such as the exponential, gamma, beta, weibull or Gumbel probability functions.

In general, the process of fitting and selecting a marginal distribution is simple and straightforward to implement. Firstly, a set of candidate distributions must be proposed to model the data set. These distributions should be fitted to the data set by means of standard techniques, such as the method of moments or the maximum likelihood method (see, e.g., [Ang and Tang, 2007](#)). Subsequently, the best fit distribution will be selected and for this purpose, different goodness of fit techniques can be employed. Among all the existing goodness of fit techniques, some widely used are worth mentioning, such as the Kolmogorov-Smirnov test, Cramér-von Mises criterion, Anderson-Darling test or the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These last two techniques, AIC and BIC, will be discussed in more detail later in this document, while the others techniques can be consulted in the standard references cited above or in other probability texts.

#### 4.4.2 Measures of correlation

Dependence is one of the most studied and important concepts in probabilistic and statistic analysis, and hence in reliability studies. Stochastic dependence has a great impact not only in the construction of a joint density function but also in the estimation of the probability of failure ([Jogdeo, 1982](#); [Nelsen, 2007](#)).

In practice, there is a wide range of degrees of dependence among geotechnical random variables. However, dependence in geotechnics is commonly characterized by a simple scalar number that varies between  $[-1, 1]$  and that evaluates the strength and direction of this property. These kind of coefficients are commonly known as measures of correlation, and although they are quite useful and easy to calculate, correlation coefficients serve only as a first approximation to the whole structure of dependence of two random variables.

There is a wide variety of correlation measures in the literature, each with its own aims, scopes, and particular properties (see, e.g., [Kotz and Drouet, 2001](#)). However, in this document, we are mainly interested in two categories, the linear correlation measures, and the measures of concordance. Linear correlation measures are important as they are the most popular in practice, and they are also the basis for elliptical distributions (a concept that will be discussed later), such as the multivariate Gaussian distribution. However, their use in non-elliptical distributions is misleading, erroneous and sometimes useless ([Dutfoy and Lebrun, 2009](#); [Embrechts et al., 2001](#)). On the other hand, the concordance measures are important since they are the most suitable correlation measures to be used in conjunction with copula theory.

The concepts of linear correlation and concordance will be discussed in more detail below. And, in particular, the Pearson's rho correlation coefficient will be introduced as a measure of linear correlation, and the Spearman's rho and Kendall's tau coefficients will be introduced as measures of concordance.

#### 4.4.2.1 Linear correlation

Linear relationship between two random variables,  $X$  and  $Y$ , can be expressed by the following equation:

$$E(Y|X = x) = \alpha + \beta x \quad (4-16)$$

where  $\alpha$  is the intercept and  $\beta$  the slope of the straight line that relate the random variables.

In this way, equation (4-16) states that by having  $X$ , one could predict the values of  $Y$  by employing a direct linear relationship. Nonetheless, two random variables are seldom perfectly associated in a linear way. Conversely, the straight line that relates the variables must be fitted to the available information, and even so the dispersion of the data may not allow a good adjustment. Thus, it is necessary to evaluate the suitability of the linear relationship between two random variables, and for this purpose the coefficients of linear correlation are employed. According to [Embrechts et al. \(2001\)](#) and [Ang and Tang \(2007\)](#), linear correlation can be defined as a measure of suitability of linear dependence. In particular, *Pearson's rho correlation coefficient* is introduced below.

**Pearson's rho** This is one of the most popular and extensively used measures of dependence, given its simplicity and its intrinsic relationship with multivariate gaussian distributions. Pearson's rho correlation coefficient is a measure of linear dependence between two random variables. This measure is expressed in terms of the covariance between the two variables and their respective variance, as follows.

$$\rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)} \sqrt{\text{var}(X_2)}} \quad (4-17)$$

Note that an extension to more than two dimensions ( $d > 2$ ) is straightforward, since the correlation is evaluated pairwise and a  $d \times d$  square correlation matrix is formed. For example, given three random variables,  $X_1$ ,  $X_2$  and  $X_3$ , the correlation matrix is expressed as:

$$\sigma = \begin{bmatrix} 1 & \rho(X_1, X_2) & \rho(X_1, X_3) \\ \rho(X_2, X_1) & 1 & \rho(X_2, X_3) \\ \rho(X_3, X_1) & \rho(X_3, X_2) & 1 \end{bmatrix}$$

It can be shown from equation (4-17) that  $\rho(X_1, X_2) \in [-1, +1]$ ; here,  $-1$  or  $+1$  would mean a perfect linear relationship between the given two variables, either positive or negative, and  $0$  would mean linear uncorrelatedness. It is worth noting that linear uncorrelatedness does not mean independence, since two variables may have no linear relationship but rather other types of relationships (see, e.g., [Ang and Tang, 2007](#)).

Nonetheless, equation (4-17) expresses that Pearson's rho correlation between two random variables depends not only on the covariance of these variables, but also on their variance, and consequently on their marginal distributions. In this sense, we can extract two important remarks.

1. The range of values that  $\rho(X_1, X_2)$  can adopt will be bounded between  $\rho_{min}$  and  $\rho_{max}$  (i.e.,  $[\rho_{min}, \rho_{max}] \subset [-1, +1]$ ), where  $\rho_{min}$  and  $\rho_{max}$  are functions of the standard deviation  $\sigma$  of the marginal distributions.
2. Pearson's correlation coefficient  $\rho$  depends on both, joint distribution and marginal distributions of the random variables. This behavior is against the nature of copula theory, where marginal distributions and dependence are isolated.

Since Pearson's correlation coefficient is only invariant in strictly increasing linear transformations, but not in strictly increasing transformations, it is necessary to explore other measures of dependence that better agree with the nature of the copulas; the measures of concordance are especially suitable for this purpose.

#### 4.4.2.2 Concordance

Before introducing the two main rank correlation coefficients, namely *Spearman's rho* and *Kendall's tau*, it is important to explore the concept of concordance.

Following [Nelsen \(2007\)](#), let us assume two independent observations  $(x_i, y_i)$  and  $(x_j, y_j)$  from a vector of continuous random variables  $(X, Y)$ . The pair of observations are said to be *concordant* if  $x_i < x_j$  whenever  $y_i < y_j$ , or if  $x_i > x_j$  whenever  $y_i > y_j$ . In the same way, these two observations are said to be *discordant* if  $x_i < x_j$  whenever  $y_i > y_j$  or if  $x_i > x_j$  whenever  $y_i < y_j$ . Alternatively, it could be formulated that two observations  $(x_i, y_i)$  and  $(x_j, y_j)$  are concordant if  $(x_i - x_j)(y_i - y_j) > 0$ , or discordant if  $(x_i - x_j)(y_i - y_j) < 0$ . In this sense, concordance is a measure of associativity between large or small values of two random variables.

*Rank correlations* are native measures of concordance for two random variables ([McNeil et al., 2015](#)). Essentially, a rank correlation is a scalar measure that is based on the ranks of the data instead of the actual numerical values. From copula theory viewpoint, rank correlation coefficients only depend on the copula and not on the marginal distributions, which contrasts with Pearson's rho correlation coefficient that depends on both. This fact makes rank correlations the natural and most appropriate measure of dependence for copulas ([McNeil et al., 2015](#)).

With this context, the two well-known nonparametric measures of dependence, namely *Spearman's rho rank correlation* and *Kendall's tau rank correlation*, can be defined as follows.

**Spearman's rho** Informally, Spearman's rho coefficient ( $\rho_s$ ) could be defined as the linear correlation of the transformed random variables. In this way, given two random variables  $X$  and  $Y$ , with CDFs  $F_X$  and  $F_Y$ , respectively, Spearman's rho coefficient is defined as

$$\rho_s = \rho(F_X(x), F_Y(y))$$

When the probability distribution of the random variables is not certainty known, it is convenient to define Spearman's rho coefficient in terms of the ranks of the data. These ranks, denoted as  $R_i$  and  $S_i$ , result from the ordering of the sample of the two random variables,  $X$  and  $Y$  (see [Ang and Tang, 2007](#), for details). Then, Spearman's rho can be defined as:

$$\rho_s(R, S) = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}$$

where

$$\bar{R} = \bar{S} = \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \sum_{i=1}^n S_i = \frac{n+1}{2}$$

**Kendall's tau** Conceptually, Kendall's tau coefficient ( $\tau$ ) is defined as the probability of concordance minus the probability of discordance

$$\tau(X, Y) = P((X - \tilde{X})(Y - \tilde{Y}) > 0) - P((X - \tilde{X})(Y - \tilde{Y}) < 0)$$

where  $\tilde{X}$  and  $\tilde{Y}$  are an independent copies of  $X$  and  $Y$ , respectively, i.e, a second version with the same distribution but independent of the first.

Analogously, a sample version of Kendall's tau coefficient is expressed in the following equation:

$$\tau = \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)} P_n - 1$$

where  $P_n$  is the number of concordant pairs and  $Q_n$  the number of discordant pairs of the sample.

Both Spearman's rho and Kendall's tau, as well as Pearson's rho correlation coefficient, are easily calculable in two dimensions. Furthermore, extensions of these coefficients to more than two dimensions are straightforward, since only pairwise evaluations of the variables are required to form a correlation matrix, as described above for the correlation matrix of Pearson's rho coefficients.

Rank measures of dependence, such as Spearman's rho and Kendall's tau, have several properties that make them especially suitable for their use in copula theory. For instance:

1. Both Spearman's rho and Kendall's tau are symmetric and take values within  $[-1, 1]$ , regardless the marginal distributions.
2. A value equal to zero in either of these two rank coefficients is a stronger indicator of independence than Pearson's rho coefficient; however, such hypothesis must be corroborated with other tests.
3. Rank correlation values of  $+1$  or  $-1$  indicate comonotonicity and countermonotonicity respectively (see, e.g., [Embrechts et al., 2002](#)).

4. Both rank correlations do not depend on marginal distributions, but rather solely on the adopted copula.
5. Both rank coefficients, unlike linear correlation, are invariant under strictly increasing transformations and are also less sensitive to outliers.

So, rank measures, either Spearman's rho or Kendall's tau, are the most suitable and natural measures to be used in the representation of the strength and direction of dependence between two random variables, in copula theory.

### 4.4.3 Popular families of copulas

A great variety of copula functions could be found in the literature, each one with its properties and dependence structure. Also, it is common to group these copulas into some families or categories, taking into account some similarities in their properties or in how these copulas are constructed.

Particularly in the geotechnical engineering framework, and after a review of the state-of-the-art, three families of copulas are mainly used, namely Elliptical, Archimedean, and Plackett copulas. These families/classes contain a wide range of copula functions, each one with its properties, advantages, and limitations to represent certain dependence structures.

In this section, we will review the aforementioned three families of copulas. Some theoretical background and explanation of their properties will be exposed. For an in deep explanation, the reader is referred for example, to [Nelsen \(2007\)](#), [Joe \(1997\)](#) or [Embrechts et al. \(2001\)](#).

There exist other families of copulas, or copulas functions, widely used and quite studied in several fields, such as the Farlie-Gumbel-Morgenstern (FGM), Ali-Mikhail-Haq (AMH) or Marshall-Olkin copulas. However, the utility of these particular copulas has been rather scarce in geotechnics, mainly due to the reduced range of degrees of dependence that they are able to handle.

#### 4.4.3.1 Elliptical copulas

Elliptical copulas stem from elliptical distributions as a consequence of the *Sklar's theorem*. The shape of these distributions is symmetric and from here stems their name, since the isocurves of their density functions are ellipse-like forms in  $\mathbb{R}^n$ .

An elliptical distribution, which is denoted as  $E_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  in  $\mathbb{R}^n$ , mainly consists of two parameters, as follows:

1. A localization vector  $\boldsymbol{\mu} \in \mathbb{R}^n$ .
2. A scatter matrix  $\boldsymbol{\Sigma} \in \mathbb{M}_{n,n}^+$ , symmetric and positive semi-definite (see, e.g., [Ang and Tang, 2007](#)).

These distributions/copulas are radially symmetric around their respective localization vector, and their scatter is represented by the scatter matrix. Another important remark about these distributions is that linear correlation coefficient is their natural measure of dependence, so for a random vector, each element  $R_{ij}$  of the correlation matrix  $\mathbf{R}$  can be defined as:

$$R_{ij} = \rho(X_i, X_j) = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.$$

Disadvantages regarded to linear correlation have already been studied before in this document. Furthermore, elliptical multivariate distributions, and the definition of their correlation matrix  $\mathbf{R}$ , require that all marginal distributions be elliptical (Embrechts et al., 2001). Nonetheless, these two pitfalls are solved using copula theory.

As in practice we may need a elliptical dependence structure with marginal distributions that are not the same and that are not necessarily elliptical, other alternatives must be explored. In this regard, robust linear correlation estimators were proposed by Embrechts et al. (2001). Furthermore, with the support of copula theory, an elliptical copula can be used in conjunction with non-necessarily elliptical marginal distributions.

In order to apply robust linear correlation estimators, rank measures of dependence have to be used. Specifically, the relation between Kendall's tau and Pearson's rho is suitable for this purpose. In this way, copula parameters of dependence of elliptical copulas, and hence their matrix of correlation, can be computed through the respective Kendall's tau coefficients between pairs of random variables. This relation is given by the following equation (Frees and Valdez, 1998):

$$\tau(X_i, X_j) = \frac{2}{\pi} \arcsin(R_{ij}) \quad (4-18)$$

where  $i$  and  $j$  are the indices of the correlation matrix  $\mathbf{R}$ . Note that if  $i = j$  then  $R_{ij} = 1$ .

With this brief introduction to elliptical distributions, we are now in a position to present two of the most used elliptical copulas in geotechnical engineering, namely the Gaussian copula and the Student- $t$  copula. Readers interested in more details about theory of elliptical distributions and elliptical copulas can find a good exposure of this topic in Embrechts et al. (2001).

Bivariate equations of both Gaussian copula and Student- $t$  copula are presented in table 4-1. Both copulas are comprehensive copulas, i.e. they both can approach to countermonotonicity  $W$  and comonotonicity  $M_d$ . Furthermore, equation(4-18) holds for these two copulas as a way of fitting them to a given Kendall's tau. Extending these two copulas to more than two-dimensions is straightforward, and this fact is one of the reasons of their popularity.

Although the copula dependence parameter of both copulas can be fitted by equation(4-18), Student- $t$  copula has an additional parameter of adjustment, namely degrees of freedom  $\nu$ . While Gaussian copula is a one-parameter copula, the Student- $t$  copula is a two-parameter copula. The number of

**Table 4-1:** Bivariate copula function  $C$  and copula density function  $c$  for Gaussian and Student- $t$  copulas

Copula	$C(u, v; \theta)$	$c(u, v; \theta)$
Gaussian	$\Phi_{\theta}(\Phi^{-1}(u), \Phi^{-1}(v))$	$\frac{1}{\sqrt{1-\theta^2}} \exp\left[-\frac{\xi_1^2\theta^2 - 2\theta\xi_1\xi_2 + \xi_2^2\theta^2}{2(1-\theta^2)}\right];$ $\xi_1 = \Phi^{-1}(u); \xi_2 = \Phi^{-1}(v)$
Student- $t$	$t_{v,\theta}(t_v^{-1}(u), t_v^{-1}(v))$	$\frac{\Gamma(\frac{v+2}{2})\Gamma(\frac{v}{2})}{\Gamma^2(\frac{v+1}{2})(1-\theta^2)^{1/2}} \frac{\left(1 + \frac{(T_v^{-1}(u))^2 - 2\theta T_v^{-1}(u)T_v^{-1}(v) + (T_v^{-1}(v))^2}{v(1-\theta^2)}\right)^{-\frac{v+2}{2}}}{\left(1 + \frac{(T_v^{-1}(u))^2}{v}\right)^{-\frac{v+1}{2}} \left(1 + \frac{(T_v^{-1}(v))^2}{v}\right)^{-\frac{v+1}{2}}}$

$T_{v+1}$  = univariate Student's  $t$  CDF with  $v + 1$  degrees of freedom

$\Gamma$  = Gamma function.

**Table 4-2:** Tail dependence of Gaussian and Student- $t$  copulas

Copula	$\lambda_l$	$\lambda_u$
Gaussian	0	0
Student- $t$	$2T_{v+1}\left(-\frac{\sqrt{v+1}\sqrt{1-\theta}}{\sqrt{1+\theta}}\right)$	$2T_{v+1}\left(-\frac{\sqrt{v+1}\sqrt{1-\theta}}{\sqrt{1+\theta}}\right)$

$T_{v+1}$  = univariate Student's  $t$  CDF with  $v + 1$  degrees of freedom

degrees of freedom of a Student- $t$  copula controls the strength of both tail dependence coefficients, i.e.  $\lambda_u$  and  $\lambda_l$ , as stated in the equations presented in table 4-2.

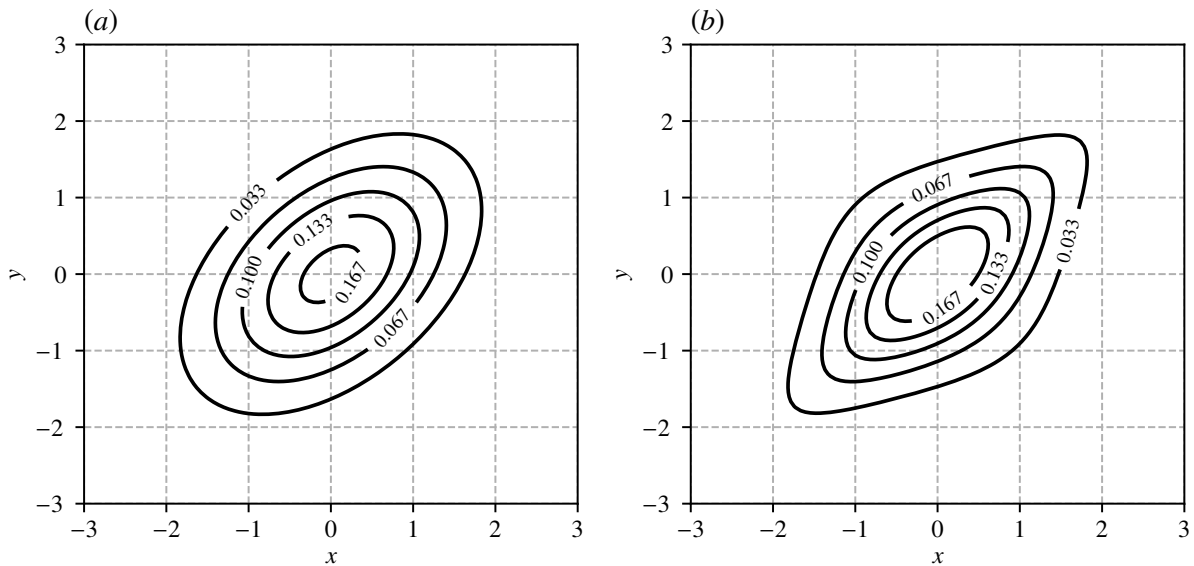
Note that as a consequence of the radial symmetry property of elliptical copulas, both coefficients of tail dependence of the Student- $t$  copula are equal. Additionally, as  $v$  increases, the tail dependence of the Student- $t$  copula becomes weaker, and vice versa, as  $v$  decreases, the tail dependence of the Student- $t$  copula becomes stronger. In this way, when  $v$  tends to infinity, both  $\lambda_u$  and  $\lambda_l$  approach to zero.

It is worth mentioning that the Gaussian copula does not have tail dependence, i.e.  $\lambda_u = \lambda_l = 0$ . However, what is truly interesting here is that the Gaussian copula is a particular case of the Student- $t$  copula. Student- $t$  copula becomes a Gaussian copula when its tail dependence is equal to zero, i.e., when  $v$  tends to infinity. Therefore, Student- $t$  copula has more general dependence structure than Gaussian copula. Some authors, such as Embrechts et al. (2001), support the use of the Student- $t$  copula as an alternative to the commonly used Gaussian dependence structure.

The number of degrees of freedom of a Student- $t$  copula can be estimated from data using fitting methods based on the concept of maximum-likelihood-estimation, like the ones presented in section 4.4.4. When no data is available and just the rank measure of dependence is known, some authors such as McNeil et al. (2015) recommend to adopt a value of 4 for the degrees of freedom.

Algorithms for modelling both Gaussian and Student- $t$  copulas can be found in Embrechts et al.

(2001) or McNeil et al. (2015). For illustrative purposes, we construct the Figure 4-4 in order to show the shape of the resulting bivariate distribution when these two copulas are employed. Figure 4-4 corresponds to two random variables that follow a standard normal distribution and that are related with a Kendall's tau coefficient of 0.3. Three degrees of freedom are used for the Student- $t$  copula.



**Figure 4-4:** Contour plots of two bivariate elliptical copulas with standard normal marginal distributions. (a) Gaussian copula ( $\tau = 0.3$ ), and (b) Student- $t$  copula ( $\tau = 0.3$  and  $\nu = 3$ )

Note in figure 4-4 that even though Gaussian copula and Student- $t$  copula are radially symmetrical and have the same dependence parameter, their shape is different thanks to the tail dependence of the Student- $t$  copula. Tail dependence will play a key role in reliability analysis, as shown in section 4.7.

#### 4.4.3.2 Archimedean copulas

Before defining Archimedean copulas, the concept of generator function must be introduced. Following Nelsen (2007) and Embrechts et al. (2001), a *generator function*  $\varphi(t)$  can be defined as a continuous strictly decreasing function that maps from  $[0, 1]$  to  $[0, \infty]$ , such that  $\varphi(1) = 0$ .

A generator function has a quasi-inverse  $\varphi^{(-1)}$  that maps from  $[0, \infty] \rightarrow [0, 1]$ , as follows:

$$\varphi^{(-1)} = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0) \\ 0, & \varphi(0) \leq t \leq \infty \end{cases}$$

Clearly,  $\varphi^{(-1)}$  is continuous and decreasing on  $[0, \infty]$ , and it is also strictly decreasing on  $[0, \varphi(0)]$ . Thus, on the interval  $[0, 1]$ ,  $\varphi^{(-1)}(\varphi(v)) = v$ . In this way:

$$\varphi(\varphi^{(-1)}(t)) = \begin{cases} t, & 0 \leq t \leq \varphi(0) \\ \varphi(0), & \varphi(0) \leq t \leq \infty \end{cases}$$

A bivariate Archimedean copula is defined in the following way:

$$C(u, v) = \varphi^{(-1)}(\varphi(u) + \varphi(v))$$

where  $C$  is a copula if and only if  $\varphi$  is convex. Note that if  $\varphi(0) = \infty$ , then  $\varphi$  is a strict generator, and thus,  $\varphi^{(-1)} = \varphi^{-1}$ , which constitutes a strict archimedean copula  $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$ .

Archimedean copulas constitute one of the most popular and flexible classes of copulas. The ease to construct these copulas, the ease to sample from them, and the wide variety of dependence structures, make this class of copulas a suitable option to model dependence in geotechnical engineering.

Although Archimedean copulas overcome some limitations of Elliptical copulas we introduce in the previous section, there are some disadvantages. Unlike Elliptical copulas, Archimedean copulas are not derived from the *Sklar's theorem*, have closed-form expression, and can model both radial and non-radial symmetry. However, despite Archimedean copulas are especially suitable for modeling dependence of two random variables, their extension to more dimensions is not as straightforward as in the case of Elliptical copulas. Actually, for extending this class of copulas to more than two dimensions some requirements must be fulfilled, and even so, there are some limitations over these kinds of procedures.

However, the complexity of the final Archimedean copula increases as the dimension grows. Additionally, these extensions are too restrictive since the same dependence structure is applied to all random variables, which may be different from reality. In this regard, some modern techniques such as vine copula theory (see section 4.2.7) overcome these types of disadvantages and are more suitable for modeling high-dimensions dependence structures when using Archimedean copulas. Readers interested in a succinct review of the extension of Archimedean copulas to more than two dimensions are referred to [Joe \(1997\)](#) and the references cited therein.

Many generator functions can be found in the literature, and hence many Archimedean copulas can be constructed. Among all Archimedean copulas, there are four widely used in geotechnical engineering for modeling dependence, namely Frank, Gumbel, Clayton, and Nelsen No 16 copulas. Among other similarities, the aforementioned four copulas have in common the capacity of modeling a wide spectrum of dependence, which is especially desirable in a copula to be used in geotechnics.

In particular, Frank copula is a comprehensive copula, Gumbel and Clayton copulas can model a complete spectrum of positive dependence, and No 16 copula can model a complete spectrum of

negative dependence and even a weak positive dependence. Additionally, there are some variants of these copulas, which extend their applicability to other structures of dependence. For example, Clayton copula can be modified to model both positive and negative dependence, survival copulas can be used or even a rotated version of the original copulas can be applied. These variants are not covered here, but readers interested in this topic are encouraged to refer to [Nelsen \(2007\)](#).

As the generator function is the keystone of Archimedean Copulas, it can be used to estimate some of the copula properties. For example, copula parameter of dependence can be estimated from Kendall's tau using the respective copula generator function, and the same for the coefficients of tail dependence. Thus, complex equations such as the ones of tail dependence (equations (4-6) and (4-7)), or the one that relates Kendall's tau and the copula parameter of dependence (see section 4.4.4) can be simplified.

When the first derivate of  $\varphi^{-1}$  is equal to minus infinity, i.e.  $\varphi^{-1'}(0) = -\infty$ , coefficients of tail dependence for archimedean copulas can be defined as:

$$\lambda_u = 2 - 2 \lim_{s \rightarrow 0} \left[ \frac{\varphi^{-1'}(2s)}{\varphi^{-1'}(s)} \right]$$

and

$$\lambda_l = 2 \lim_{s \rightarrow \infty} \left[ \frac{\varphi^{-1'}(2s)}{\varphi^{-1'}(s)} \right]$$

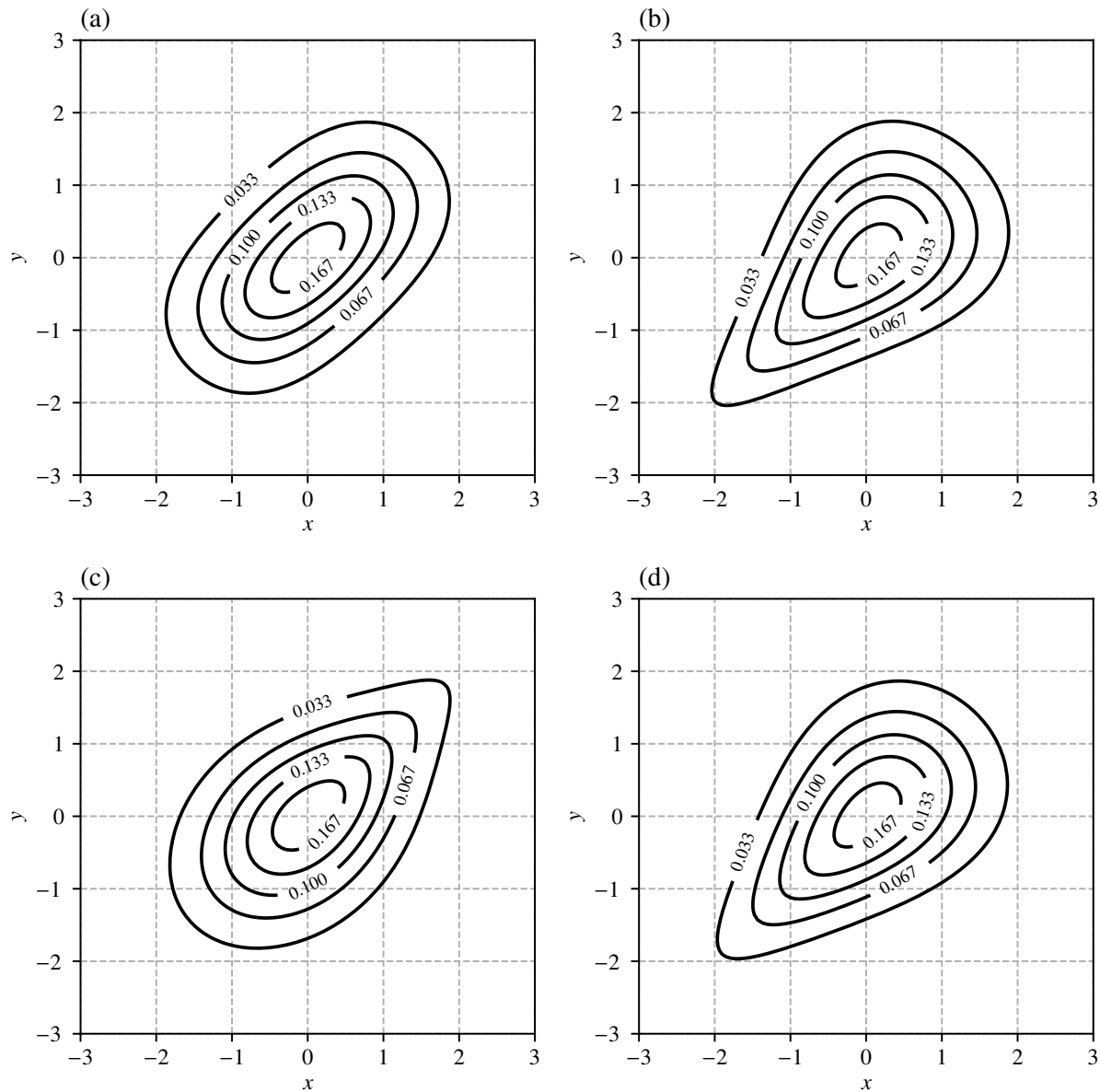
Conversely, if  $\varphi^{-1'}(0)$  is finite, then both coefficients of tail dependence will be equal to zero. Proofs of these formulas are exposed by [Genest and MacKay \(1986\)](#), [Joe \(1997\)](#) and [Embrechts et al. \(2001\)](#).

Table 4-3 presents some useful information about the most common Archimedean copulas in geotechnical engineering, namely Frank, Gumbel, Clayton, and Nelsen No. 16 copulas. Furthermore, there are several algorithms reported in the literature for sampling from different archimedean copulas. A good exposure of these algorithms can be found in [Nelsen \(2007\)](#), [Embrechts et al. \(2001\)](#) or [Phoon and Ching \(2014\)](#).

Coefficients of tail dependence of the studied Archimedean copulas are summarized in table 4-4.

For comparative purposes, contour plots of the bivariate PDFs generated through these four copulas are presented in figure 4-5. A coefficient of Kendall's tau  $\tau = 0.3$  is used to calibrate these copulas, which is the same used for calibrating Gaussian and Student-*t* copulas of figure 4-4. Furthermore, the same marginal distributions, i.e. standard normal, are also for constructing figure 4-5.

Note that the contour plots of the different copulas in figure 4-5 differ considerably, and these in turn differ from those of figure 4-4. These differences will be reflected in the probabilities of failure for any performance function, as we will see in section 4.7.



**Figure 4-5:** Contour plots of four bivariate archimedean copulas with standard normal marginal distributions. (a) Frank copula ( $\tau = 0.3$ ), (b) Nelsen No 16 copula ( $\tau = 0.3$ ), (c) Gumbel copula ( $\tau = 0.3$ ), and (d) Clayton copula ( $\tau = 0.3$ )

**Table 4-3:** Generator function  $\varphi$ , copula function  $C$ , and copula density function  $c$  in a bivariate case for some popular Archimedean copulas used in geotechnics.

Copula	$C(u, v; \theta)$	$c(u, v; \theta)$	$\varphi(t)$
Frank	$-\frac{1}{\theta} \ln \left[ 1 + \frac{(e^{-u\theta} - 1)(e^{-v\theta} - 1)}{e^{-\theta} - 1} \right]$	$\frac{-\theta(e^{-\theta} - 1)e^{-\theta(u+v)}}{[(e^{-\theta} - 1) + (e^{-u\theta} - 1)(e^{-v\theta} - 1)]^2}$	$-\ln \left[ \frac{e^{-t\theta} - 1}{e^{-\theta} - 1} \right]$
Gumbel	$\exp \left[ -\left( (-\ln u)^\theta + (-\ln v)^\theta \right)^{1/\theta} \right]$	$\frac{(-\ln u)^{\theta-1} (-\ln v)^{\theta-1} \left[ s^{\frac{2-2\theta}{\theta}} - (1-\theta) s^{\frac{1-2\theta}{\theta}} \right]}{uve^{s^{1/\theta}}}$ ; $s = (-\ln u)^\theta + (-\ln v)^\theta$	$(-\ln t)^\theta$
Clayton	$[u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}$	$\frac{(1+\theta)(uv)^{-(\theta+1)}}{(u^{-\theta} + v^{-\theta} - 1)^{\frac{1+2\theta}{\theta}}}$	$\frac{1}{\theta} (t^{-\theta} - 1)$
No 16	$\frac{1}{2} (s + \sqrt{s^2 + 4\theta})$ ; $s = u + v - 1 - \theta \left( \frac{1}{u} + \frac{1}{v} - 1 \right)$	$\frac{1}{2} \left( 1 + \frac{\theta}{u^2} \right) \left( 1 + \frac{\theta}{v^2} \right) z^{-1/2} \dots$ $\dots \times \left\{ -\frac{1}{z} \left[ u + v - 1 - \theta \left( \frac{1}{u} + \frac{1}{v} - 1 \right) \right]^2 - 1 \right\}$ ; $z = \left[ u + v - 1 - \theta \left( \frac{1}{u} + \frac{1}{v} - 1 \right) \right]^2 + 4\theta$	$\left( \frac{\theta}{\tau} + 1 \right) (1 - t)$

**Table 4-4:** Coefficients of tail dependence for some popular Archimedean copulas used in geotechnics.

Copula	$\lambda_l$	$\lambda_u$
Frank	0	0
Gumbel	0	$2 - 2^{1/\theta}$
Clayton	$2^{-1/\theta}$	0
No 16	0.5	0

**4.4.3.3 Plackett copula**

Plackett copula is another widely applied copula in geotechnical engineering. This copula belongs to a family of its own, and it is a particular example of copulas constructed through algebraic methods. As the theory of the algebraic methods for constructing copulas is out of the scope of this document, readers are encouraged to refer to [Nelsen \(2007\)](#), where a good introduction to this topic is presented.

A bivariate Plackett copula has the advantage of being a comprehensive copula, which is quite useful for modeling the dependence of geotechnical random variables. Additionally, among other properties, Plackett copula does not have tail dependence, either positive or negative, and is radially-symmetric, just like the Gaussian or Student-*t* copulas.

However, the extension of Plackett copula to more than two dimensions is not as simple as in the case of elliptical copulas, and this fact constitutes one of its disadvantages. Although some researches such as [Zhang and Singh \(2019\)](#) have presented an extension of Plackett copula to

three-dimensions, its complexity increases significantly, like its limitations. In this regard, it is worth mentioning again that vine copula theory is especially helpful for modeling dependence in high-dimensions, and we recommend its use whenever necessary.

Additionally, there is no closed-form expression in the relation between rank coefficients and the Plackett copula parameter of dependence. Therefore, when adjusting a Plackett copula to a dataset, the parameter of the copula must be approximated through numerical methods. More of this adjustment will be explained in section 4.4.4.

Table 4-5 presents the equations of the bivariate Plackett copula and its associated PDF.

**Table 4-5:** Bivariate copula function  $C$  and copula density function  $c$  for the Plackett copula.

Copula	$C(u, v; \theta)$	$c(u, v; \theta)$
Plackett	$\frac{s - \sqrt{s^2 - 4uv\theta(\theta-1)}}{2(\theta-1)}$ $s = 1 + (\theta - 1)(u + v)$	$\frac{\theta[1 + (\theta-1)(u+v-2uv)]}{\{[1 + (\theta-1)(u+v)]^2 - 4uv\theta(\theta-1)\}^{3/2}}$

Coefficients of tail dependence of the Plackett copula are summarized in table 4-6.

**Table 4-6:** Coefficients of tail dependence for the Plackett copula

Copula	$\lambda_l$	$\lambda_u$
Plackett	0	0

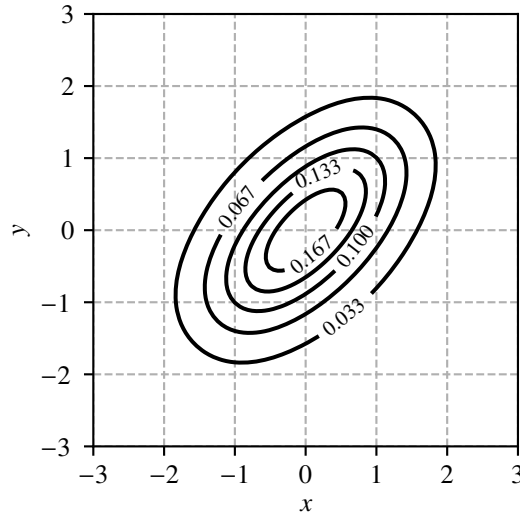
Algorithms for sampling from a bivariate Plackett copula can be found in [Nelsen \(2007\)](#) or [Li et al. \(2013\)](#). Additionally, as it has been done with the other copulas of the different families, the contour plot of the bivariate PDF generated from of a Plackett copula are presented. For comparative purposes, figure 4-6 is constructed using a Kendall's tau coefficient equal to 0.3 for two random variables distributed as standard normal probability functions.

#### 4.4.4 Estimation of the parameters of a copula

A copula may be specified in terms of one or several parameters that are gathered in a vector  $\theta$ . In practice,  $\theta$  is adjusted to a dataset using one of many available methods. In the following, an introduction to these methods is given, after [McNeil et al. \(2015\)](#), [Genest and Favre \(2007\)](#), and [Durrleman et al. \(2000\)](#). In particular, the method of moments and the maximum likelihood (with two of its variants: the IFM and the CML) will be explained.

##### 4.4.4.1 Method of moments based on rank correlations

As stated in section 4.4.2, rank correlation statistics, like Kendall's tau or Spearman's rho, are the natural measures of dependence in copula theory. Even, it is possible to relate both Kendall's tau



**Figure 4-6:** Contour plots of a bivariate Plackett copula for  $\tau = 0.3$  with standard normal marginal distributions.

or Spearman's rho to the parameter of dependence of a copula.

When a copula  $C$  is absolutely continuous, then Kendall's tau correlation coefficient  $\tau$ , between two random variables  $X$  and  $Y$ , is defined using the Stieltjes integral (Schweizer et al., 1981; Nelsen, 2007):

$$\tau(X, Y) = \tau(u, v) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \quad (4-19)$$

In the same way, Spearman's rho correlation coefficient  $\rho_s$  between two random variables  $X$  and  $Y$  can be defined in terms of an absolutely continuous copula  $C$ , as (Schweizer et al., 1981):

$$\rho_s(X, Y) = \tau(u, v) = 12 \int_0^1 \int_0^1 (C(u, v) - uv) du dv \quad (4-20)$$

Observe that a reverse procedure is also possible in equations (4-19) and (4-20). That is to say, from a given Kendall's tau or Spearman's rho, the dependence parameter of a copula could be estimated. However, it is not usual to have the exact value of the rank measure but rather an approximation so, in this case, the dependence parameter defined through equation (4-19) or (4-20) is also an approximation.

In some cases, there exist simplifications of equations (4-19) and (4-20). Nonetheless, unfortunately in some other cases, these simplifications are not possible, and even equations (4-19) and (4-20) may not have an analytical solution.

In the case of Archimedean copulas equation (4-19) can be simplified as a one-dimensional integral

in terms of the generator function and its derivate, as (Genest and MacKay, 1986):

$$\tau = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt \quad (4-21)$$

Furthermore, for some archimedean copulas, such as Gumbel and Clayton copulas, equation (4-21) can be further simplified, obtaining a quite simple algebraic expression (see table 4-7). Nonetheless, not all Archimedean copulas achieve this level of simplification, such as in the case of Ali-Mikhail-Haq copula (see, e.g., Nelsen, 2007), so numerical methods have to be employed over equation (4-21) or equation (4-19).

Regarding elliptical copulas, Pearson's rho correlation coefficient is their main descriptor of dependence, i.e. their copula parameter of dependence. However, as stated in section 4.4.2, this coefficient has many disadvantages from the copula's theory viewpoint. To overcome these disadvantages, relationships between Pearson's rho and rank measures of dependence can be used, such as the ones of equations (4-19) or (4-20).

These estimates lead to the already presented equation (4-18), that relates Kendall's tau and the parameter of dependence of elliptical copulas. Additionally, there also exists a direct estimation of the elliptical copula parameter of dependence from Spearman's rho coefficient, which is given by (McNeil et al., 2015):

$$\rho_s(X_i, X_j) = \frac{6}{\pi} \arcsin\left(\frac{R_{ij}}{2}\right) \quad (4-22)$$

where  $R_{ij}$  is the  $ij$ -th element of the correlation matrix that contains the copula parameters of dependence.

However, it is worth saying that equation (4-22) results in good approximations for Gaussian copulas, but this is not the case of other elliptical copulas such as the Student- $t$  (Hult and Lindskog, 2002). In this sense, it is preferable to employ the relationship  $\tau - \rho$ , given by the equation (4-18), since it is more robust and applicable to all elliptical copulas (McNeil et al., 2015).

As stated before, not all copulas have a closed-form solution of equations (4-19) or (4-20), such as the Plackett copula. In these cases, numerical methods have to be employed to approximate the copula parameter of dependence by employing rank measures.

Table 4-7 summarizes the simplified equations that relate  $\tau$  and the copula parameter  $\theta$  for the most commonly used copula functions in geotechnics. When the relationship is not specified, it means that equation (4-19) must be solved using numerical methods. Additionally, table 4-7 also presents information on the range of  $\tau$  values that the studied copulas in section 4.4.3 can model, as well as the range of values that their respective dependence parameter can adopt.

#### 4.4.4.2 Maximum Likelihood Estimator

A well-known alternative to the method of moments is the method of maximum likelihood estimation (MLE), which may be more robust and suitable than the method of moments when the

**Table 4-7:** Simplified relationships between  $\tau$  and  $\theta$  for some widely employed copulas in geotechnics.

Copula	Range of $\tau$	Relationship $\tau - \theta$	Range of $\theta$
Gaussian	$[-1, 1]$	$\tau = \frac{2}{\pi} \sin^{-1}(\theta)$	$[-1, 1]$
Student- $t$	$[-1, 1]$	$\tau = \frac{2}{\pi} \sin^{-1}(\theta)$	$[-1, 1]$
Plackett	$[-1, 1]$	–	$(0, \infty) \setminus \{1\}$
Frank*	$[-1, 1] \setminus \{0\}$	$\tau = 1 - \frac{4}{\delta} + 4 \frac{D_1(\delta)}{\delta}$	$(-\infty, \infty) \setminus \{0\}$
Gumbel	$[0, 1]$	$\tau = 1 - \frac{1}{\theta}$	$[1, \infty)$
Clayton	$[-1, 1] \setminus \{0\}$	$\tau = \frac{\theta}{2+\theta}$	$[-1, \infty) \setminus \{0\}$
No. 16	$[-1, 1/3)$	–	$[0, \infty)$

\*Debye function  $D_1(\delta) = \int_0^\delta \frac{x/\delta}{e^x - 1} dx$

information is scarce (Ang and Tang, 2007; Genest and Favre, 2007). This method allows to estimate parameters of a probability distribution by maximizing its likelihood function based on the available data.

In the framework of copula theory, the MLE maximizes concurrently both marginal distributions parameters and copula parameters. In other words, by a single optimization, the most probable points in the parameter space for both structures are obtained.

Following Durrleman et al. (2000), consider a continuous case for both the copula function  $C$  and all marginal distributions  $F_{X_i}$ , where  $i = 1, 2, \dots, d$ , and  $d$  stands for the number of random variables. Now, assuming that the copula density  $c$  exists, and hence the joint PDF is given by equation (4-3), the expression of the log-likelihood can be written as (Joe and Xu, 1996):

$$l(\theta) = \sum_{t=1}^N \ln c(F_1(x_1^t; \theta_1), F_2(x_2^t; \theta_2) \dots F_d(x_d^t; \theta_d); \alpha) + \sum_{t=1}^N \sum_{i=1}^d \ln f_i(x_i^t; \theta_i) \quad (4-23)$$

where  $x_i^t$  is the  $d$ -dimensional sample of size  $N$ ,  $\theta_n$  and  $\alpha$  are the vectors of parameters of the marginal distributions and the copula function, respectively, and  $\theta = (\theta_1, \theta_2, \dots, \theta_d; \alpha)$ .

By maximizing equation (4-23), the parameter vector  $\theta$  is obtained, which corresponds to the maximum likelihood estimative. That is to say,  $\theta$  is the parameter vector, from both copula functions and marginal distributions, that maximizes the likelihood function of the model.

It is worth mentioning that the availability of the data is mandatory to carry out the MLE method since otherwise there is no way to maximize equation (4-23). By contrast, the method of moments based on rank correlations can be carried out by using the data or directly by a known Kendall's tau / Spearman's rho. Furthermore, note that both MLE method and moments method are strongly

influenced by the amount and quality of data, since the greater amount/quality of data the more accurate estimates are obtained.

In some cases, the MLE method may seem less attractive than the method of moments, since it may involve a higher computational calculation cost over the copula density function and marginal distributions at the same time. However, this should not be a problem considering current computational capabilities.

On the other hand, the applicability of the MLE method is greater than the one of the method of moments, since the MLE method is especially useful when the copula has more than one fitting-parameter (Genest and Favre, 2007). In this way, the degrees of freedom of a Student- $t$  copula can be adjusted using the MLE, which is not possible using the method of moments. Thus, the MLE is a more comprehensive strategy than the method of moments, since the latter only adjusts the copula parameter of dependence and needs the help of other methods to adjust the rest of the parameters, in case they exist.

#### 4.4.4.3 The inference functions from margins and the canonical maximum likelihood methods

Although the MLE is a quite robust and widely applicable method, it may become unfeasible in very high dimensions. The MLE method estimates both parameters of the marginal distributions and parameters of the copula function at the same time, and therefore it may be too computationally expensive. In this way, following the nature of the maximum likelihood methodology, it is necessary to explore other methodologies that are computationally more efficient but at the same time more robust and widely applicable.

The *inference functions for margins method* (IFM), appears as an alternative for estimating both copula function parameters and marginal distributions parameters (Joe and Xu, 1996). The IFM method differs from the MLE method in the sense that the optimization is not done at the same time for the copula function and the marginal distributions. Instead, this procedure splits the MLE method into two stages. The first one fits the parameters of the univariate marginals distributions  $\hat{\theta}_i$  using

$$\hat{\theta}_i = \arg \max_{\theta} \sum_{t=1}^N \ln f_i(x_t^i; \theta_i)$$

and in the second one, the copula vector of parameters  $\alpha$  is computed based on  $\hat{\theta}_i$ , as follows

$$\hat{\alpha} = \arg \max_{\alpha} \sum_{t=1}^N \ln c \left( F_1(x_1^t; \hat{\theta}_1), F_2(x_2^t; \hat{\theta}_2), \dots, F_d(x_d^t; \hat{\theta}_d); \alpha \right)$$

As we can see, the IFM method follows the same nature of the MLE method, so they both share the same advantages over the method of moments. However, the IFM method is less computationally expensive than the MLE method, so its use in practice is sometimes preferred.

Additionally, following the idea of maximum likelihood estimation, there is a third technique known as the *canonical maximum likelihood* (CML) method. In this method, copula parameters are estimated without specifying marginal distributions, and for this purpose, empirical distributions are used to transform the original data  $\{(x_1^t, \dots, x_d^t)\}_{t=1}^n$  into values in the uniform space  $\{(u_1^t, \dots, u_d^t)\}_{t=1}^n$ . In this sense, we can rewrite equation (4-23) as follows:

$$\hat{\alpha} = \arg \max_{\alpha} \sum_{t=1}^N \ln c(\hat{u}_1^t, \hat{u}_2^t, \dots, \hat{u}_d^t; \alpha)$$

where  $\hat{u}_n^t$  is obtained through the empirical marginal distribution as follows:

$$\hat{u}_n^t = \frac{R_t}{N + 1} \quad (4-24)$$

in which  $R_t$  is the rank of the sample  $t$  and  $N$  is the total number of samples.

As a remark, these three maximum likelihood based methods (i.e. MLE, IFM and CML) result in quantities quite similar but with some degree of discrepancy. Especially, differences are more pronounced between the CML method and the other two methods, since the MLE and IFM methods do take into account a parametric function for the marginal distributions and the CML method does not. According to [Durrleman et al. \(2000\)](#), if there exists a considerable difference between the results of the CML method and the ones of the MLE-IFM methods, these discrepancies may be indicative that marginal distributions are not well defined. In this case, the criteria for adjusting/selecting marginal distributions must be checked.

#### 4.4.5 Tests of goodness-of-fit for copulas

Let us consider a sample  $\vartheta = \{x_i \in \mathbb{R}^d : i = 1, 2, \dots, N\}$ , and a set of copulas  $C$  fitted to that sample through any method exposed in section 4.4.4. The question that arises is which copula in  $C$  provides the best representation of dependence for the sample? In this regard, it is necessary to apply some criteria to evaluate which copula among the candidate copulas  $C$  is the most suitable.

In the case of marginal distributions, there are several popular goodness-of-fit tests, such as the Chi-Square ( $\chi^2$ ), the Kolmogorov-Smirnov (K-S), or the Anderson-Darling (A-D) tests (see, e.g., [Ang and Tang, 2007](#), for more details). However, in the case of copula functions, there are no widely used and accepted methods such as in the selection of the marginal distributions. Attention has been commonly focused on how to fit copulas to data (as seen in section 4.4.4) and not on how to choose the most appropriate copula for a data set. For this reason, the selection of the best fit copula, especially when the information is scarce, is an open question in the state-of-the-art ([Genest et al., 2009](#)).

Following [Genest et al. \(2009\)](#), methods that seek to select the most appropriate copula for representing the dependence of a  $d$ -dimensional sample can be divided into four categories:

- Special procedures for a certain family of copulas, like in the case of archimedean or Gaussian copulas (see, e.g., [Malevergne et al., 2003](#); [Genest and Rivest, 1993](#); [Barbe et al., 1996](#)).
- Procedures applicable to any type of copula, but whose implementation require an arbitrary parameter, kernel-weight functions or ad hoc categorization of the data in order to apply an analogous of the chi-square test.
- Procedures applicable to any type of copula, in which it is not required the adoption of any parameter or special strategy. These methods are called *blanket tests*, and their application is based on some transformations such as the Kendall's transform, the Rosenblatt's transform or the empirical copula (see [Genest et al., 2009](#)).
- Standard goodness-of-fit tests, whose implementation is aimed at choosing the most appropriate copula in a set of candidate copulas  $C$ , but not to test whether the selected model is suitable in the light of some measure such as the  $P$ -value (see, e.g., [Ang and Tang, 2007](#); [Kottegoda and Rosso, 2008](#), for its definition).

During the review of this state-of-the-art, we found that the most used tests for selecting the best copula, and almost the only ones employed, belong to this last category. Thus, standard goodness-of-fit tests, specifically the Akaike-Information-Criterion (AIC) and the Bayesian-Information-Criterion (BIC), are widely used for selecting the better copula in  $C$  when representing the dependence of geotechnical variables. However, these two tests do not validate the choice of the copula.

To clarify this concept, take into account that standard goodness-of-fit-tests seek to define the copula, within a set of candidate copulas  $C$ , that best represents the available information. However, if all copulas in  $C$  are poor representatives of the sample dependence, the least poor representation of dependence (i.e., copula) will be selected. Therefore, all copulas in  $C$  must be responsibly proposed, taking into account some characteristics such as the range of dependence values that the copula can handle, tail dependence, symmetry, etc. For example, if two random variables have a negative correlation,  $C$  must contain copulas that can model negative dependence, and not copulas that are only restricted to positive dependence.

Other methods worth delving are the blanket tests, since these validate the suitability of a copula through some criteria such as the  $P$ -value. An excellent exposure of these methods is given by [Genest and Favre \(2007\)](#) and [Genest et al. \(2009\)](#), so readers are encouraged to consult these references. In this document we will introduce two blanket tests based on the empirical copula and the statistics the Cramér-von Mises and Kolmogorov-Smirnov tests.

In this way, section [4.4.5.1](#) is devoted to two goodness-of-fit tests based on the empirical copula, specifically the Cramér-von Mises and Kolmogorov-Smirnov goodness-of-fit tests. On the other hand, section [4.4.5.2](#) is devoted to two standard goodness-of-fit tests, specifically the Akaike Information Criterion and the Bayesian Information Criterion.

#### 4.4.5.1 Goodness-of-fit test based on Empirical Copula

Starting from the concept of empirical copula already introduced in section 4.2.6, it is possible to apply analogous methodologies to the univariate Cramér-von Mises and Kolmogorov-Smirnov goodness-of-fit tests in order to evaluate the suitability of a copula function when modeling the dependence of a data sample. For this purpose, it is necessary to compare the distance between the empirical copula  $C_n$  and the parametric copula  $C_\theta$  under the null hypothesis  $H_0$ , i.e., the evaluated copula model cannot be rejected.

In this way, Cramér-von Mises test statistic  $S_n$ , employing an empirical copula, is defined as:

$$S_n = \int_{[0,1]^d} \sqrt{N} (C_n(\mathbf{u}) - C_\theta(\mathbf{u})) dC_n(\mathbf{u}) \quad (4-25)$$

Analogously, Kolmogorov-Smirnov test statistic ( $T_n$ ) can be expressed as:

$$T_n = \sup_{\mathbf{u} \in [0,1]^d} \left| \sqrt{N} (C_n(\mathbf{u}) - C_\theta(\mathbf{u})) \right|. \quad (4-26)$$

In the above equations,  $N$  is the sample size and  $C_n(\mathbf{u})$  is the empirical copula constructed through equation (4-10).

When  $C_\theta$  has an analytical expression, then  $S_n$  and  $T_n$  can be directly computed using equations (4-25) and (4-26), respectively; otherwise, a Monte Carlo simulation with  $m$  trials would have to be applied:

1. Draw a  $d$ -dimensional random sample  $\mathbf{U}_1^*, \dots, \mathbf{U}_m^*$  from  $C_\theta$ , with  $m \geq n$ .
2. Approximate  $\hat{C}_{\theta_n}$  by:

$$B_m^*(\mathbf{u}) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(\mathbf{U}_i^* \leq \mathbf{u}), \quad \mathbf{u} \in [0, 1]^d$$

3. Compute  $S_n$  and/or  $T_n$  by:

$$S_n = \sum_{i=1}^n \{C_n(\mathbf{U}_i) - B_m^*(\mathbf{U}_i)\}^2 \quad (4-27)$$

$$T_n = \sup_{\mathbf{u} \in [0,1]^d} \left| \sqrt{n} (C_n(\mathbf{U}_i) - B_m^*(\mathbf{U}_i)) \right| \quad (4-28)$$

After fitting the copula function, the  $P$ -value can be approximated from both tests using a parametric bootstrap simulation conducted  $N$  times, being  $N$  some larger integer, as follows:

1. Draw a random sample  $\mathbf{X}_{i,k}^*, \dots, \mathbf{X}_{n,k}^*$  from the copula  $C_{\theta_n}$ , and compute the respective rank vectors  $\mathbf{R}_{i,k}^*, \dots, \mathbf{R}_{n,k}^*$ , using equation (4-9).

2. Compute  $U_{i,k}^x$ , for  $i = 1, \dots, n$ , using equation (4-8).

3. Let

$$C_n^*(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{U}_i^* \leq \mathbf{u})$$

4. For the evaluated copula, estimate the fitting parameters from  $\mathbf{U}_i^*$ . This can be done using any of the methods described in section 4.4.4.

5. Compute the tests statistic either directly through equations (4-25) and (4-26), or approximately through equations (4-27) and (4-28).

6. Approximate the  $P$ -value of the tests statistic employing any of the following equations:

$$P_{value} = \frac{1}{N} \sum_{k=1}^N \mathbb{I}(S_{n,k}^* > S_n)$$

$$P_{value} = \frac{1}{N} \sum_{k=1}^N \mathbb{I}(T_{n,k}^* > T_n)$$

#### 4.4.5.2 Akaike and Bayesian information criterion

The Akaike Information Criterion (Akaike, 1974), and the Bayesian Information Criterion (Schwarz et al., 1978), hereinafter referred to as AIC and BIC correspondingly, are criteria of model selection within a finite set of candidate models. That is to say, both AIC and BIC evaluate the quality of a model and compare it with the quality of the other candidate models that represent the data set. In practice, these two criteria are closely related, although their origin and nature is different (Burnham and Anderson, 2004; Vrieze, 2012).

Akaike information criterion is defined as:

$$AIC = 2k - 2l \quad (4-29)$$

meanwhile Bayesian information criterion is given by:

$$BIC = k \ln N - 2 \ln l, \quad (4-30)$$

here  $k$  represents number of model/copula fitting parameters (for a parametric copula where  $\theta$  is a one-dimensional vector, then  $k = 1$ ),  $l$  stands for the value of the likelihood function of the copula density function adjuted with  $\theta$ ;  $N$  is the sample size, equivalent to the number of observations or measures.

In both cases, the copula  $C$  with the smallest AIC/BIC in the set  $C$  must be the preferred copula. Equations (4-29) and (4-30) can be divided into two parts, one that corresponds to the adjustment

of the model to the available data, represented by the likelihood function  $L$ , and the other corresponding to the number of parameters used by the model, represented by  $k$ . Thus, both tests have a preference for the models that best fit data, but in turn, penalize the complexity of the model given by the number of adjusted parameters. Therefore, the both tests look for models that neither overfit nor underfit the true model.

Although these two tests are quite similar, and even the AIC can be derived from a Bayesian framework just like the BIC (Burnham and Anderson, 2004), there are several differences between them. For example, penalty for the number of parameters is greater in the case of the BIC test in comparison to the AIC test. Additionally, some authors argue that the use of the AIC or BIC tests must be evaluated depending of the target task (Burnham and Anderson, 2002, 2004).

Particularly, the BIC test is argued as especially useful for selecting the “true model”. Thus, if the true model is in the set of candidate models, the BIC test will select it with a probability of one when the sample size tends to infinity. Nonetheless, in geotechnical engineering practice, there is no true or perfect model since these are just approximations of reality, and as an aggravating factor, sample sizes are usually small. Therefore, under these circumstances, the AIC has shown a better performance than the BIC.

Nonetheless, as the formulas of AIC and BIC tests are easily calculable, it is a good practice to compute both and compare the selected best model within the candidate models set. Usually, the best model will be the same independently of the used test. A detailed comparison between these two methods is exposed by Burnham and Anderson (2002, 2004); Vrieze (2012).

## 4.5 Some applications of copula theory in geotechnical engineering

This section is devoted to present several uses of copula theory in geotechnical engineering, in concordance with the current state-of-the-art. Section 4.5.1 describes how they have been employed to model the dependence between shear strength parameters. Section 4.5.2 discusses the applicability of copulas for modeling settlement parameters. The uses of copulas in geotechnical earthquake engineering and soil spatial interpolation are presented in section 4.5.3 and section 4.5.4, respectively. Other uses of copulas in relation to geotechnics are presented in section 4.5.5. Since reliability analysis is one of the main applications of copulas in geotechnics, section 4.7 will be devoted to discuss this topic.

### 4.5.1 Shear strength parameters of the Mohr-Coulomb failure criterion

In geotechnical engineering, Mohr-Coulomb failure criterion is a popular mathematical model that describes the response of soil and rock materials to shear and normal stresses. This model

is represented by a linear envelope that relates the shear strength of the material and the applied normal stress, as:

$$\tau = \sigma \tan \phi + c \quad (4-31)$$

where  $\tau$  is the shear strength,  $\sigma$  is the normal stress, and  $c$  and  $\phi$  are the shear strength parameters, cohesion and friction angle, respectively. Note that in equation (4-31), cohesion is the intercept of the envelope with the  $\tau$  axis, and  $\tan \phi$  is the slope of the envelope. Furthermore, depending on the loading conditions both  $c$  and  $\phi$  can be either total stress parameters or effective stress parameters.

Both cohesion and inner friction angle are commonly obtained from the same laboratory test, either by a direct shear test or by a triaxial test (see [Lambe and Whitman, 1991](#); [Mitchell et al., 2005](#), for details of these two tests and their modalities). For this reason, and since both parameters come from the same soil or rock sample, there exists an intrinsic relation between them that can be studied from a probabilistic perspective. Some researches have pointed out a negative dependence between cohesion and friction angle (see, e.g., [Low, 2007](#); [Li et al., 2011](#); [Tang et al., 2012](#)), and others a positive one (see, e.g., [Lumb, 1970](#); [Wolff, 1985](#)). Whatever the case, this dependence exists and cannot be neglected, as stated by [Lumb \(1970\)](#), [Matsuo and Kuroda \(1974\)](#), [Cherubini \(2000\)](#), [Fenton and Griffiths \(2003\)](#) or [Fellin and Oberguggenberger \(2012\)](#). Nonetheless, some researches have worked with cohesion and friction angle as independent parameters (see, e.g., [Alonso, 1976](#); [Tobutt and Richards, 1979](#); [Li and Lumb, 1987](#); [Lee and Chi, 2011](#); [Nguyen and Chowdhury, 1984](#)), although this is an unwarranted assumption.

A rudimentary approximation to the dependence between cohesion and inner friction angle is expressed in terms of Pearson's rho correlation coefficient (employed by, e.g., [Lizarraga and Lai, 2014](#); [Parker et al., 2008](#); [Hata et al., 2008](#)). However, this approximation is not adequate considering the large dispersion between values of these two parameters and the disadvantages of Pearson's rho correlation coefficient that have already been exposed in section 4.4.2. Nonetheless, since Pearson's rho coefficient is quite simple to compute and is the basis of the popular multivariate normal distribution, it has been widely employed in geotechnical engineering (see [Baecher and Christian, 2005](#); [Soubra and Mao, 2012](#), for further details).

Given that the multivariate normal distribution is quite restrictive and in general has several disadvantages (see section 4.3), other alternatives had to be explored to model the dependence between soil shear strength parameters. The Nataf isoprobabilistic transformation was one of these alternatives, and in particular, this transformation was one of the most widely used approaches for modeling dependence before the development of copula theory. However, as exposed in section 4.4.2, the Nataf transformation is indeed a particular case of the Gaussian copula. Thus, in light of copula theory, soil shear strength parameters have been modeled employing only one dependence structure, the one given by the bivariate Gaussian copula.

To fill this gap, some researches have evaluated other alternatives to the Gaussian copula to model the dependence between shear strength parameters, such as the Plackett, Frank, or Nelsen No. 16

copulas (see, e.g., [Tang et al., 2015](#); [Xu et al., 2016a](#); [Tang et al., 2017](#); [Wu, 2013a,b](#)). Note that the aforementioned copulas can handle a wide spectrum of negative and positive dependence. This property is needed since the dependence between soil shear strength parameters varies in a wide spectrum of negative values, and even sometimes positive values are reported in literature. Some popular copulas such as the Farlie-Gumbel-Morgenstern (FGM) or the Ali-Mikhail-Haq (AMH), which are restricted to a relatively weak dependence ([Nelsen, 2007](#); [Sriboonchitta and Kreinovich, 2018](#)), may not be suitable to model the dependence of soil and rock shear strength parameters.

Furthermore, it is worth mentioning that the copulas in the set of candidate copulas  $C$ , for representing a dataset of soil shear strength parameters, must be varied and have different properties such as asymmetry, tail dependence, range of dependence values, etc. These differences allow to evaluate several possibilities in order to select the one that best fits the structure of dependence of the dataset.

As an illustrative example, in this section we are carrying out the exercise of defining a copula model for an effective shear strength dataset. This exercise will be executed based on the procedure studied in section 4.4 and it is in concordance with the current state-of-the-art on the construction of copula functions for soil/rock shear strength parameters. Comparisons with other investigations of the same nature, and some remarks, that in the judgment of the authors are worth highlighting, are also presented. It is worth mentioning that this exercise can be extended to other geotechnical parameters, so it is not limited to soil/rock shear strength parameters.

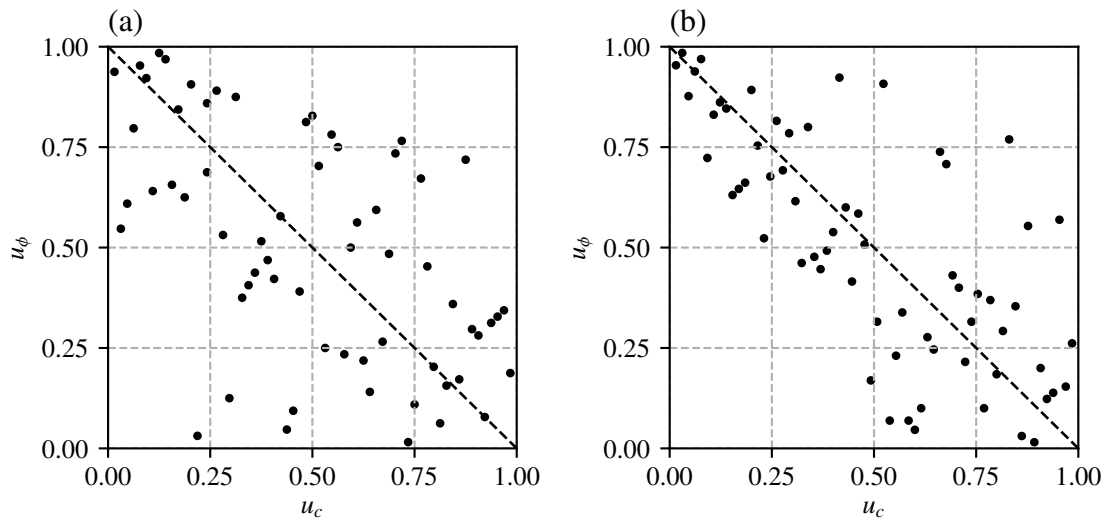
For this example, the effective shear strength dataset from the Xiaolangdi Hydropower Station is employed (readers can consult this dataset in [Zhang et al., 2013](#)). In particular, the measures obtained from two loading conditions of triaxial tests, namely consolidated-drained test (CD) and consolidated-undrained test (CU), are used. The sample size of each one of these datasets is equal to 63 and 64 respectively.

This dataset is particularly useful since it has already been studied from a probabilistic perspective, and even from a copula theory framework, in other works (see, e.g., [Phoon and Ching, 2014](#); [Li et al., 2015b](#)). A comprehensive soil description and a probabilistic description of the data can be found in [Yu \(2006\)](#) and [Zhang et al. \(2013\)](#).

As a preliminary step, it is necessary to study the type of dependence between cohesion and friction angle of each triaxial test modality. For this purpose, dependence can be isolated of the effect of marginal distributions by applying the empirical distribution over the original dataset.

Following this guideline, figure 4-7 contains two scatter plots in the uniform space for each loading condition of triaxial test. These scatter plots are obtained after applying the transformation given by the empirical distribution, equation (4-24).

Scatter plots of figure 4-7 show that there exists a negative trend between cohesion and friction angle for both modalities of triaxial tests. Points of all the scatter plots of the figure 4-7 are located next to a  $-45^\circ$  slope line, which represents a perfect negative dependence.



**Figure 4-7:** Scatter plots in the standard uniform space of  $c$  vs  $\phi$ , from the dataset of the Xiaolangdi Hydropower Station after employing the transformation given by the empirical distribution. (a) CD dataset, (b) CU dataset.

The negative correlation between cohesion and friction angle for both triaxial test modality is validated through the respective Pearson's rho and Kendall's tau coefficients. The resulting Pearson's rho coefficients for the CD test and CU test are equal to  $-0.544$  and  $-0.702$ , respectively. The resulting Kendall's tau coefficients for the CD test and CU test are equal to  $-0.384$  and  $-0.544$ , respectively.

Considering this negative dependence, one might already have some candidate copulas in mind to model the dataset. For instance, the Gaussian, Plackett, Frank and Nelsen No. 16 would be good options. Furthermore, some other copulas that have not been extensively used, so far, to model shear strength parameters can be proposed. For instance, the Student- $t$  Copula is a suitable option, although it has been only used in a few works of copulas in geotechnics (e.g., Wu, 2015, 2013a). Some researchers point out that the Student- $t$  copula is more general than the Gaussian copula, and should be used instead of this last one when possible (see, e.g., Embrechts et al., 2001).

It is worth mentioning that some other authors have even worked with copulas that are not suitable to model negative dependence in a first instance, such as the Gumbel or Clayton copulas (e.g. Wu, 2015; Xu et al., 2016a). For this purpose, they transform one of the random variables to force a positive dependence. This operation is also valid and allows to take advantage of the particular properties of interest of some copulas. Details of this transformation can be found in the above references or in Nelsen (2007).

Now, the next step is to adjust the candidate copulas to the available data. Recalling section 4.4.4, there exist four practical methods to adjust the candidate copulas to the available information. The

MLE method calibrates both copula function and marginal distributions at the same time, while the IFM, CML, and moment methods, calibrate marginal distributions and the copula function separately.

In the literature, these four fitting methods have been indiscriminately used for adjusting copulas to shear strength parameters. The MLE method was employed by [Zhang et al. \(2014\)](#) to adjust marginal distributions and copulas at the same time. On the other hand, [Huang et al. \(2014\)](#), [Phoon and Ching \(2014\)](#), and [Wu \(2015\)](#) employed the IFM method to firstly adjust marginal distributions and posteriorly copula functions. The CML was exposed in the work of [Tang et al. \(2017\)](#). Finally, the method of moments has been used by [Xu et al. \(2016a\)](#), [Wu \(2013a\)](#), or [Tang et al. \(2013\)](#).

It is possible to conclude that the most used methods for fitting copula functions to datasets of shear strength parameters correspond to the IFM method and the method of moments. Note that unlike the IFM method, the method of moments only estimates the copula parameter of dependence, so other copula parameters and marginal distributions must be adjusted independently.

However, the importance of fitting copula functions and marginal distributions at the same time, i.e., of using the MLE method to adjust copulas to data, was pointed out by [Zhang et al. \(2014\)](#). In this regard, [Zhang et al. \(2014\)](#) expresses that copulas affect the fitting values of the marginal distributions, so for a robust estimation of the models, the MLE method must be used instead of the other methods.

We agree with [Zhang et al. \(2014\)](#) at considering the MLE method as the most robust method for fitting copula functions to data; some other researches, such as [Durrleman et al. \(2000\)](#), support this postulate. This consideration is based on the fact that data is obtained as samples and not as populations, so there will be discrepancies among the results of the four fitting methods. The scarcer the information, the greater the discrepancies between results of the methods; and vice versa, the more information, the smaller the differences among results of the methods. The MLE method, by fitting marginal distributions and copula at the same time, maximizes the likelihood function of the model and may lead to the best model estimative based on the available data; this in comparison to other methods that adjust separately marginal distributions and copulas.

Nonetheless, when fitting copula models to data, [Zhang et al. \(2014\)](#) allows negative values of soil cohesion, which is physically impossible. In this way, the work of [Zhang et al. \(2014\)](#), and its conclusions, must be reviewed in light of this physical restriction. It is important to be aware of the physical restrictions of the random variables, since they has a direct impact on the final results. Thus, further studies should be carried out to quantify the discrepancies between the fit values of models, obtained through all the different methods used for this purpose.

As an example in this document, several marginal distributions will be adjusted to the data maximizing their likelihood function. These marginal functions are the truncated normal below zero, lognormal, gamma, beta, and weibull distributions. Then, the suitability of each marginal distribu-

tions will be compared through goodness-of-fit-test.

There are several goodness-of-fit tests for univariate distributions (see e.g. [Ang and Tang, 2007](#); [Kottogoda and Rosso, 2008](#)), and some of these have been widely applied to model the distributions of the shear strength parameters. In the copula framework for modeling shear strength parameters, the Kolmogorov-Smirnov test has been used by [Huang et al. \(2014\)](#) or [Xu et al. \(2016a\)](#), and the Anderson-Darling was used by [Wu \(2013a,b\)](#). Even the AIC and BIC methods have been used as means of model selection of univariate distributions (e.g. [Phoon and Ching, 2014](#); [Tang et al., 2017](#)), although these do not constitute a formal goodness-of-fit test (see section 4.4.5.2).

In most cases, the Kolmogorov-Smirnov test is enough to assess the feasibility of a univariate probability model, and no other tests are needed. Thus, the K-S test is applied in this document to define the best-fit marginal distributions at a significance level of 0.05 ( $\alpha = 0.05$ ). Results are summarized in table 4-8.

**Table 4-8:** K-S values of the different marginals distributions. Here  $D_{max}$  stands for K-S test statistic.

Triaxial test	Parameter	truncated normal	lognormal	gamma	beta	weibull	$D_{max}$
CD	c	0.072	0.127	0.102	<b>0.06</b>	0.067	0.171
	$\phi$	0.109	<b>0.093</b>	0.098	0.146	0.151	0.171
CU	c	<b>0.056</b>	0.173	0.127	0.092	0.076	0.17
	$\phi$	<b>0.082</b>	0.111	0.1	0.103	0.091	0.17

**Note:** Bold value in each row indicates the best-fit distribution according to the K-S test.

On the other hand, AIC/BIC values of all the marginal distributions were computed and used in the original work of [Phoon and Ching \(2014, chap. 2\)](#) to define the best fitting model to data. As a comparative exercise, it is interesting to contrast best-fit marginal distributions obtained through these two methodologies, i.e. K-S test and AIC/BIC test. Table 4-9 presents the AIC/BIC values for all the parameters and marginal distributions.

**Table 4-9:** AIC/BIC values of different marginals distributions

Triaxial	Parameter	truncated normal		lognormal		gamma		beta		weibull	
		AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
CD	c	568.6	572.89	589.77	594.06	579.83	584.11	<b>567.93</b>	<b>572.22</b>	568.94	573.23
	$\phi$	300.86	305.14	300.26	304.55	<b>300.15</b>	<b>304.44</b>	307.77	312.05	310	314.29
CU	c	<b>613.14</b>	<b>617.46</b>	645.97	650.28	627.48	631.79	617.58	621.9	617.01	621.33
	$\phi$	<b>340.34</b>	<b>344.66</b>	342.11	346.43	341.03	345.35	341.75	346.07	343.85	348.16

**Note:** Bold values in each row indicate the preferred distribution according to the AIC and BIC tests.

In the first instance, it is evident that there are some minor discrepancies between some AIC/BIC values reported in this document compared with those reported by [Phoon and Ching \(2014, chap. 2\)](#).

These discrepancies are mainly due to two reasons: the fitting method and the limits of the truncated functions. While the method of moments was used by [Phoon and Ching \(2014, chap. 2\)](#), in this document the likelihood of each distribution was maximized. Therefore, it is expected that different fitting parameters were obtained, although close.

As an interesting remark, note that AIC/BIC values reported in this document are in all cases smaller than those reported by [Phoon and Ching \(2014, chap. 2\)](#). This fact indicates that maximizing the likelihood functions lead to better-fit models than those obtained by the method of moments used by [Phoon and Ching \(2014, chap. 2\)](#).

Regarding best-fit marginal distributions, both methodologies agree on their definition in the majority of cases. Note that in this particular example, most suitable marginal distributions differ in the case of inner friction angle of the CD triaxial test. Nonetheless, in this case the best-fit distributions obtained through both methodologies have quite similar values of AIC/BIC or K-S, so their data representation is similar and their differences are minor, in practice.

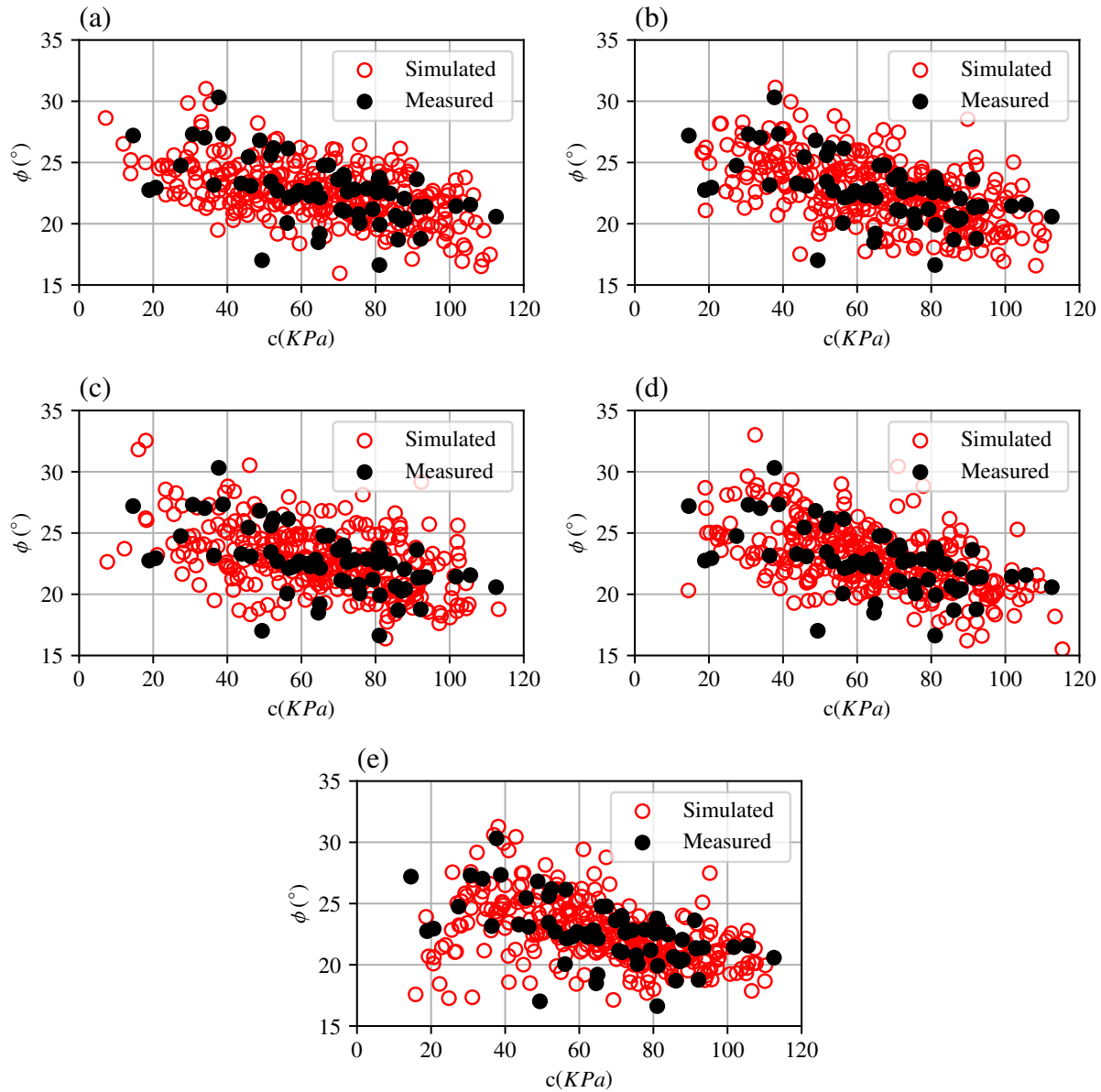
Additionally, the present work includes two marginal distributions that were not evaluated in the study of [Phoon and Ching \(2014, chap. 2\)](#), namely the gamma and beta probability distributions. Particularly, these two distributions are found suitable to model both shear strength parameters of the consolidated drained triaxial test. This find is a motivation to include different types of univariate distributions within the set of candidate distributions to represent cohesion and friction angle, or other geotechnical parameters.

After defining marginal distributions, the next step is proposing a set of copulas, adjust them to data, and select the most suitable one. For this purpose, the same copulas proposed by [Phoon and Ching \(2014, chap. 2\)](#) are evaluated in this document, i.e. Gaussian, Plackett, Frank, and Nelsen No. 16 copulas. Additionally, Student- $t$  copula is also evaluated since it is a more general copula than the Gaussian copula, and its use is recommended by some researches such as [Embrechts et al. \(2001\)](#).

The parameter of the dependence of each copula is fitted through the method of moments, i.e., using equation (4-19) or its simplifications. Furthermore, the number of degrees of freedom of the Student- $t$  copula is assumed to equal to 4.0 as recommended when it is not possible to estimate it ([McNeil et al., 2015](#)).

Although in this case a maximum likelihood estimation can be carried out to compute the number of degrees of freedom of the Student- $t$  copula, for illustrative purposes  $\nu$  is assumed in order to represent tail dependence and compare the Student- $t$  copula with the Gaussian copula.

Figure 4-8 presents scatter plots for the CD triaxial test of the Xiaolangdi Hydropower Station, in conjunction with 300 points simulated through the aforementioned 5 copulas, after their respective adjustment to the data. It is worth mentioning that for this simulation the best-fit marginal distributions defined in table 4-8 were employed.



**Figure 4-8:** Simulation of CD data from different copulas. Cohesion modeled as a beta distribution and friction angle modeled as a lognormal distribution (a) Gaussian copula, (b) Student- $t$  copula, (c) Plackett copula, (d) Frank copula, (e) Nelsen No 16 copula

It seems that all copulas well-fit to data, but evidently their suitability must be validated through goodness-of-fit tests. For this purpose, the AIC or BIC values were computed for each copula and triaxial test. The results of these calculations are summarized in table 4-10.

**Table 4-10:** AIC and BIC values for all the studied copulas in the example of shear strength parameters

Triaxial Test	Gaussian		Student- <i>t</i>		Plackett		Frank		Nelsen No. 16	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
CD	-19.65	-17.51	-13.72	-9.44	-18.05	-15.91	<b>-20.22</b>	<b>-18.07</b>	-15.4	-13.26
CU	-43.62	-41.47	-39.72	-35.4	-43.12	-40.96	<b>-44.22</b>	<b>-42.07</b>	-32.46	-30.3

**Note:** Bold values in each row indicate the preferred Copula according to the AIC and BIC tests.

From table 4-10, Frank copula is defined as the most suitable copula for the datasets of CD and CU triaxial tests. This means that the Gaussian copula is not the best representation of dependence between soil shear strength parameters of the Xiaolangdi Hydropower Station.

Gaussian dependence structure is commonly adopted without validation in practice, but in the light of copula theory, this is just one of several dependence structures. Therefore, the suitability of the Gaussian copula for modeling dependence of a dataset must be evaluated and contrasted with other copulas. The same conclusion is found in the many studies of dependence between soil/rock shear strength parameters from the framework of copula theory.

Additionally, note that in this particular case Gaussian copula performs better than Student-*t* copula, as observed in the AIC/BIC values of table 4-10. These differences between the AIC/BIC values of the Gaussian and Student-*t* copula are due to the assumed degrees of freedom. In this way, shear strength parameters from the Xiaolangdi Hydropower Station are better represented with no tail dependence. This is corroborated with the best fit copula since the coefficient of tail dependence is equal to zero for Frank copula.

There exist other shear strength datasets that have been studied from a copula theory perspective. For instance, it is worth mentioning the datasets of the Ankang Hydropower station (see Tang et al., 2013; Zhang et al., 2014), the dataset of the Jiangkou, Shuikou, and Ertan hydropower stations (see Tang et al., 2013), the dataset of the Shicuan hydropower station (see Huang et al., 2014), a shear strength dataset of soils from Shanghai (see Chao and Lin-de, 1998; Zhang et al., 2013), among other. Nonetheless, similar results and conclusions were obtained in all these works.

Furthermore, Wu (2015) conducted a study, from copula theory perspective, in which several dataset of shear strength parameters were employed. For all these datasets, Wu (2015) constructed copulas and found similar results and conclusions to those obtained in here. In general, there is no universal accepted copula for modeling shear strength parameters of soil and rocks. Each problem has its particular conditions, so copulas must be fitted to data and the best fit copula selected through the methods explained here in order to obtain the best possible representation of

dependence.

Finally, it is worth mentioning that procedures described in this section are based on the current state-of-the-art. So, these procedures could be replicated to model any dataset of shear strength parameters or even any bivariate dependence in geotechnical engineering.

### 4.5.2 Settlement

In the framework of settlement analysis, copula theory has been mainly employed for the study of model uncertainty of different equations that are used for serviceability-limit-state designs. Most of these equations relate, for different types of foundations, the applied load and the consequent displacement. To this end, these relations are commonly expressed in terms of a normalized load  $Q$  and a normalized displacement  $\eta$ . Commonly, the normalized load is obtained through a load of reference, and the normalized displacement is obtained through some geometrical factor. After the normalization, the final scatter in the curves is significantly reduced, as shown by [Dithinde et al. \(2011\)](#), [Huffman et al. \(2015\)](#) or [Uzielli and Mayne \(2011, 2012\)](#). In this regard, copulas have been used to model the final scatter of these curves, represented by the fitting parameters of their equations (i.e., model uncertainty).

Foundation load-displacement relationships can be evaluated through several equations reported in literature. In particular, two equations have been mostly employed, which are known as the hyperbolic equation and the power law equation. Hyperbolic equation can be expressed in normalized terms as:

$$Q = \frac{\eta}{a + b\eta} \quad (4-32)$$

while the power law equation can be expressed in normalized terms as:

$$Q = c\eta^d \quad (4-33)$$

where  $Q$  is the normalized load,  $\eta$  represents the normalized displacement, and  $a$ ,  $b$ ,  $c$ , and  $d$ , are model fitting parameters.

To normalize load-settlement curves several methodologies have been applied. In the case of the induced settlement, this is mostly normalized by a foundation geometry factor, commonly known as  $B$ . The parameter  $B$  can represent the diameter of the foundation, the width of the foundation, or even the equivalent footing diameter, i.e. the diameter that produces the same area as that of a square footing (see, e.g., [Mayne and Poulos, 1999](#)). The normalized settlement is hence computed as  $\eta = s/B$ , where  $s$  is the measured settlement and  $B$  the geometry normalization factor.

On the other hand, normalized load  $Q$  employs a reference capacity  $q_{ref}$  to normalize the applied load  $q$ . Reference capacity is not necessarily the true foundation capacity, but rather a known capacity, preferably obtainable through a deflection criterion ([Huffman et al., 2015](#)).

For example, following Briaud (2007), in-situ cone net tip resistance  $q_{c,net}$ , from the cone penetration test (CPT), was employed by Uzielli and Mayne (2011, 2012) as  $q_{ref}$  for normalizing applied loads on shallow footings on sand. On the other hand, following Phoon et al. (2003, 2006), the slope tangent capacity  $Q_{STC}$  was employed by Li et al. (2012, 2013) as  $q_{ref}$  for normalizing the applied load on static piles, or by Huffman et al. (2015) over spread footings on clays. Details of the cone penetration tests (CPT) and in-situ cone net tip resistance  $q_{c,net}$  are found in Robertson and Cabal (2015) and Briaud (2007). More details about the slope tangent capacity  $Q_{STC}$  are found in Phoon et al. (2003, 2006).

It is worth mentioning that a normalized displacement is not always employed in equations (4-32) or (4-33) before modeling the fitting parameter of the curves through copulas. For example, only the applied load was normalized by Dithinde et al. (2011) and Li et al. (2012, 2013) when employing equation (4-32) for representing load-displacement curves of static piles. In this case,  $\eta$  in equations (4-32) and (4-33) is replaced by the crude settlement measure  $s$ . Conversely, unlike normalized displacement, normalized load is always employed in equations (4-32) or (4-33) before modeling fitting parameters through copulas.

Copula theory has been employed for modeling the remaining scatter after normalizing the load-displacement curves dataset. In this way, dependence between parameters  $a$  and  $b$  from equation (4-32), or  $c$  and  $d$  from equation (4-33), has been studied under copula theory framework.

For instance, Uzielli and Mayne (2011) used a load-displacement database of 30 shallow footings in different sands to calibrate equation (4-33). After the calibration, they proposed a methodology to estimate the design load for different threshold settlements in shallow footings on sands, based on the CPT and for a given reliability level.

It was demonstrated by Uzielli and Mayne (2011) that the parameters  $c$  and  $d$  of equation (4-33) are dependent on each other and independent of other variables such as  $B$  or  $q_{app}$ . They employed a lognormal distribution to model both  $c$  and  $d$  parameters, and different copula functions were evaluated to represent their dependence. Frank copula was found to be the best-fit copula for  $c$  and  $d$ , and additionally, other variables such as foundation width  $B$ ,  $q_{net}$  or  $q_{app}$  were also adopted as random variables but without dependence for the analysis. Thus, a probabilistic approach was presented by Uzielli and Mayne (2011) to design vertically loaded shallow footings on sand based on the CPTu test, and taking into account model uncertainty of equation (4-33) with the support of copula theory.

Later, Uzielli and Mayne (2012) delve into the importance of taking into account the dependence between fitting parameters of the power law model (equation (4-33)). The same load-displacement database of Uzielli and Mayne (2011) was employed by Uzielli and Mayne (2012) for calibrating equation (4-33). They evaluated three procedures for modeling load-displacement curves of shallow footings on sand, namely the fully probabilistic MP, the parametric uncertainties PO, and the deterministic DT approaches. MP approach consists of a fully probabilistic method, where  $B$ ,  $q_{app}$ , and  $q_{net}$  parameters are modeled as independent random variables; additionally, both  $c$  and  $d$

parameters from equation (4-33) are modeled as random variables through a Frank copula and beta marginal distributions. The PO method comprises in a probabilistic method where  $B$ ,  $q_{app}$  and  $q_{net}$  are independent random variables, but  $c$  and  $d$  in equation (4-33) are taken as deterministic values. Finally, DT approach involves a fully deterministic method where  $B$ ,  $q_{app}$  and  $q_{net}$ , just like the parameters  $c$  and  $d$ , are taken as deterministic values.

In this way, [Uzielli and Mayne \(2012\)](#) highlights the importance of probabilistic analysis when modeling load-displacement curves; in particular, the importance of the dependence between  $c$  and  $d$  parameters of equation (4-33) is highlighted. They pointed out that the MP approach showed to be the most appropriate one, and neglecting the inherent uncertainty and dependence will result in inaccuracies. Settlement is overestimated in low loading levels and underestimated in high loading levels when the MP methodology is not employed.

Similar works were conducted by [Huffman and Stuedlein \(2014\)](#) and [Huffman et al. \(2015\)](#). The former studied, from a reliability-based point of view, the limit state design of spread footings on aggregate pier reinforced clay. The latter, studied, again from a reliability-based point of view, the immediate settlement of spread footings on clays. These two works used copulas to model the dependence between two power-law parameters, i.e.  $c$  and  $d$  from equation (4-33), in order to incorporate uncertainties into the methodologies of study. We do not delve into these works, since the use of copulas is quite similar to that employed by [Uzielli and Mayne \(2011, 2012\)](#). Readers are referred to [Uzielli and Mayne \(2011\)](#), [Uzielli and Mayne \(2012\)](#), [Huffman and Stuedlein \(2014\)](#) and [Huffman et al. \(2015\)](#) for further information.

On the other hand, several investigations that employ copula theory have been applied to the pile load-displacement database collected, organized, and categorized by [Dithinde et al. \(2011\)](#). The data was collected from various piling companies that operate in the Southern Africa countries, and it was grouped into four subsets of pile load tests, as follows: driven piles in non-cohesive soils (D-NC), bored piles in non-cohesive soils (B-NC), driven piles in cohesive soils (D-C), and bored piles in cohesive soils (B-C). The sample size of each category of pile load test is equal to 28, 30, 59 and 53, respectively. In addition to the load-displacement data, the properties of the soil and the geometric characteristics of the foundations are reported.

[Dithinde et al. \(2011\)](#) demonstrated that the scatter in the original load-displacement curves can be substantially reduced when the applied load is normalized by the interpreted capacity of the pile (see [Phoon et al., 2006](#)). This procedure leads to normalized curves that can be correctly fitted through the hyperbolic model. However, it is worth noting that they do not normalize settlement measures, so equation (4-32) is rewritten as:

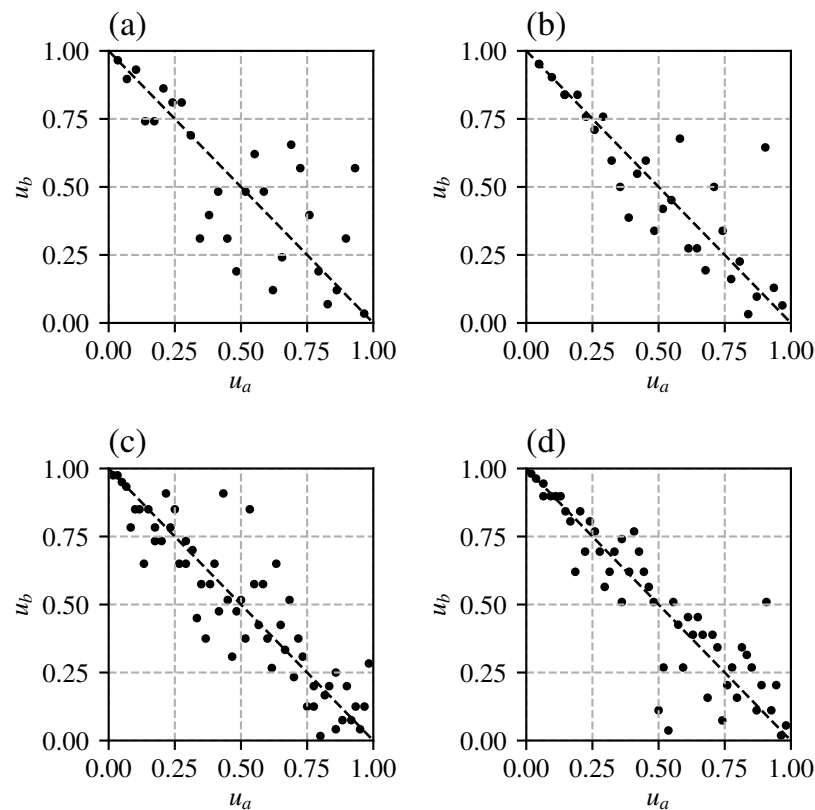
$$Q = \frac{s}{a + bs} \quad (4-34)$$

where  $Q$  is the normalized load,  $s$  the pile top displacement, and  $a$  and  $b$  are curve fitting parameters.

Curve fitting parameters from equation (4-34),  $a$  and  $b$  can be interpreted as the reciprocals of the initial slope and asymptotic value of the normalized curves. The procedure to obtain the interpreted capacity  $Q_i$  is detailed in the work of [Dithinde et al. \(2011\)](#), where a conjugation of Chin's extrapolation procedure ([Chin, 1970](#)) and Davisson's failure criterion ([Davison, 1972](#)) was employed.

After normalizing the applied load of the datasets, there is a remaining model uncertainty that can be represented modeling both curve fitting parameters. Particularly,  $a$  and  $b$  parameters of equation (4-34) are found to be mutually dependent and independent of other variables.

Figure 4-9 presents scatter plots in the standard uniform space, for each category of piles, after applying the transformation given by the empirical distribution over the pile curve-fitting-parameters. There is an evident negative trend between the two curve fitting parameters for all categories of pile load tests. Thus, copulas that model negative dependences have to be employed.



**Figure 4-9:** Scatter plots of pile curve fitting parameters from the dataset of [Dithinde et al. \(2011\)](#) in the standard uniform space after applying the transformation given by the empirical distribution. (a) D-NC, (b) B-NC, (c) D-C, (d) B-C

[Dithinde et al. \(2011\)](#) demonstrated the inherent dependence between  $a$  and  $b$  parameters of equa-

tion (4-34), and employed a simple translation lognormal model to represent it, following the procedures described by Phoon et al. (2006, 2007) or Phoon (2008, chapter 10). On the other hand, Li et al. (2012) modeled the dependence between  $a$  and  $b$  using a Gaussian copula. Later, Li et al. (2013) conducted a comparison between not only two types of Gaussian copulas (the one adjusted to data through a direct computation of the Pearson's rho coefficient and the one fitted to data through Kendall's tau coefficient) but also with other three copulas capable of handling a wide range of negative dependence, namely Plackett, Frank, and Nelsen No. 16. They used a methodology to adjust copulas analogous to the one employed to fit copulas for soil shear strength parameters of section 4.5.1.

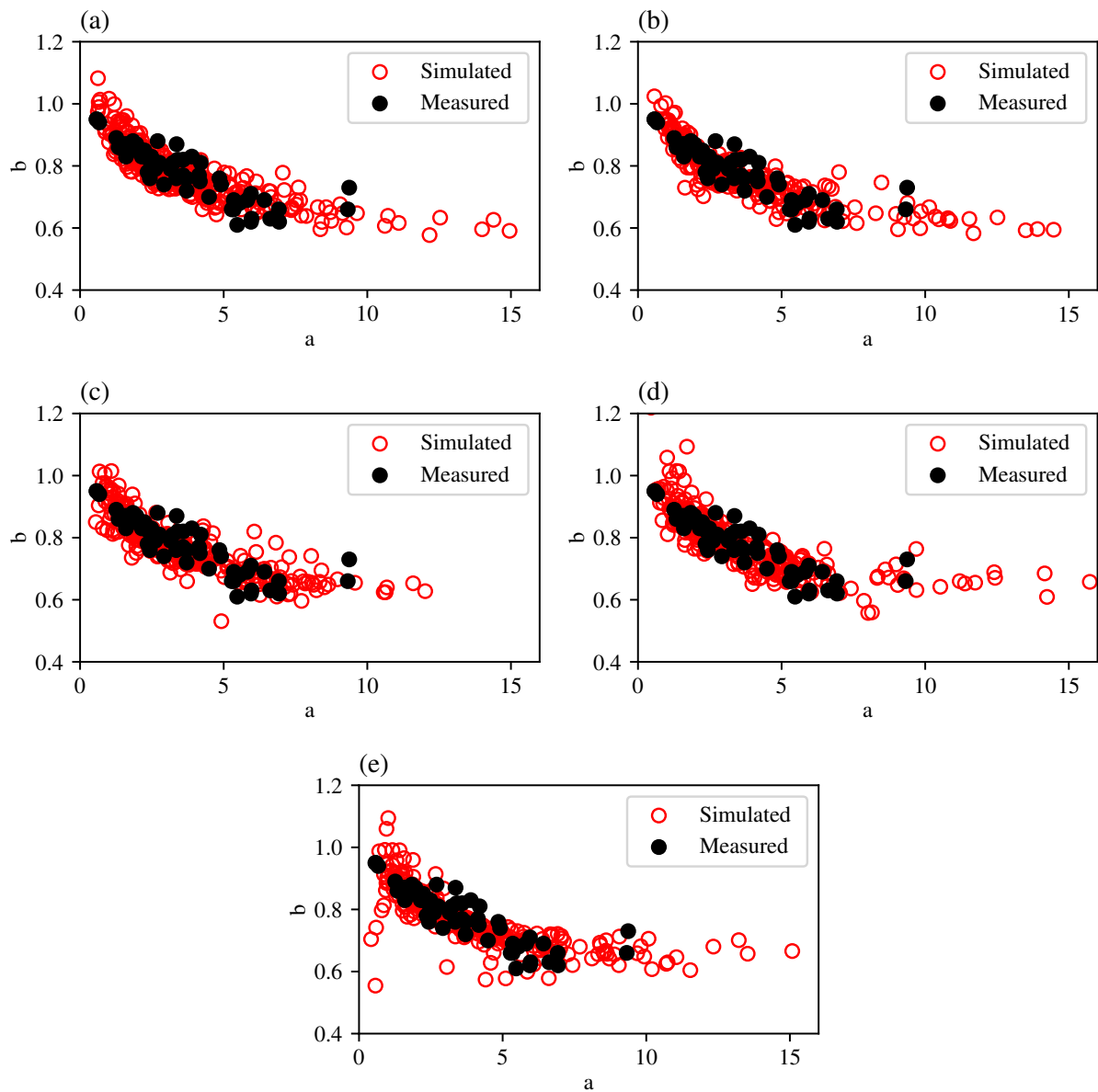
In this section an analogous modeling to that of Li et al. (2013) is performed in order to present some worth mentioning results and remarks, obtained from the dependence of the dataset of Dithinde et al. (2011). In addition to the original four families of copulas proposed by Li et al. (2013), this work also includes the Student- $t$  copula in the analysis, following the recommendations of Embrechts et al. (2001). Concerning the fitting parameters of the Gaussian copula, only the Gaussian copula based on Kendall's tau will be taken into account, and not the one based on a direct computation of Pearson's rho. As stated by Li et al. (2012, 2013), the Gaussian copula adjusted by Kendall's tau coefficient performs better than the Gaussian copula fitted by a directly computed from data Pearson's rho coefficient.

Table 4-11 presents the fitting parameters of the five different studied copulas. All the dependence parameters are defined through the method of moments by relating Kendall's tau and the copula parameter of dependence. When there is no analytical solution of the relation between Kendall's tau and the copula parameter of dependence, such as in the case of the Plackett copula, a numerical approximation is applied over equation (4-19). On the other hand, the second parameter of the Student- $t$  copula corresponds to the number of degrees of freedom, which is defined through a maximum likelihood procedure using the available data.

**Table 4-11:** Fitting parameters of the studied copulas for the pile curve fitting parameters.

Pile Test	Gaussian $\theta$	Student- $t$ $\theta; \nu$	Plackett $\theta$	Frank $\theta$	Nelsen No 16 $\theta$
D-NC	-0.806	-0.806; 1.97	0.048	-7.85	0.008
B-NC	-0.924	-0.924; 1.50	0.015	-14.139	0.002
D-C	-0.918	-0.918; 3.50	0.016	-13.512	0.003
B-C	-0.927	-0.927; 2.29	0.014	-14.471	0.002

A graphical contrast between the measured and the simulated data from all these copulas, for the driven piles in cohesive soils dataset, is shown in figure 4-10. Following Li et al. (2012, 2013), a lognormal marginal distributions are adopted for representing both  $a$  and  $b$  parameter. Similar scatter plots can be obtained for the other pile load test categories.



**Figure 4-10:** Simulation of the D-C dataset from different copulas using lognormal marginal distributions. (a) Gaussian copula, (b) Student- $t$  copula, (c) Plackett copula, (d) Frank copula, (e) Nelsen No 16 copula

As stated in section 4.5.1, it is impossible to define the best-fit copula model by a simple visual inspection of the simulated data, such as in the scatter plots of figure 4-10. For instance, the best-fit copulas must be defined quantitatively based on some criteria. For example, Li et al. (2012, 2013) employed the widely applied AIC/BIC methodologies to determine the best-fit models. This methodology is also adopted in this document. Then, AIC and BIC coefficients are computed and presented in table 4-12 for all the studied copulas and all the pile tests categories.

**Table 4-12:** AIC and BIC values for the different studied copulas in the example of pile curve fitting parameters.

Pile Test	Gaussian		Student- <i>t</i>		Plackett		Frank		Nelsen No 16	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
D-NC	-26.07	<b>-24.73</b>	<b>-26.33</b>	-23.67	-23.32	-21.99	-22.03	-20.70	-16.28	-14.95
B-NC	-42.00	-40.60	<b>-50.60</b>	-47.80	-49.98	<b>-48.58</b>	-42.99	-41.59	-44.23	-42.83
D-C	-81.64	-79.56	-86.93	-82.78	<b>-90.27</b>	<b>-88.20</b>	-90.04	-87.96	-83.62	-81.54
B-C	-73.94	-71.97	-87.83	-83.89	<b>-88.97</b>	<b>-87.00</b>	-81.79	-79.81	-84.83	-82.86

**Note:** Bold values in each row indicate the preferred Copula according to the AIC and BIC tests.

From table 4-12, there can be extracted some important observations worth mentioning. Firstly, note that although the dataset is the same, there are some minor discrepancies between the AIC/BIC values reported in this document and the ones reported by Li et al. (2013). These minor discrepancies are due to the number of decimals employed for the computation of the AIC/BIC values. Nonetheless, these minor discrepancies do not have a relevant impact on the final results since the same best-fit copulas are obtained.

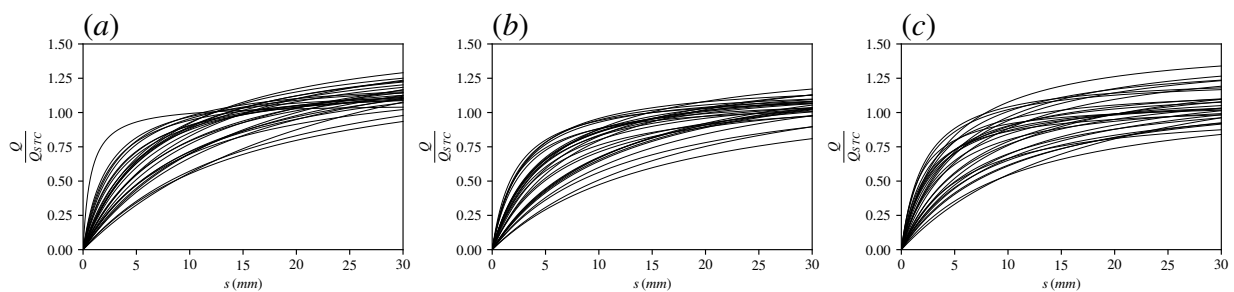
Second, observe in table 4-12 that AIC/BIC values in bold highlight the best-fit copulas. Although both selection criteria mostly agree in the determination of the best-fit copula, there are some cases where they disagree. Particularly, this situation is evident in the copula definition of the D-NC and B-NC datasets, and most specifically concerning the Student-*t* copula.

While the AIC defines the Student-*t* copula as the best-fit one for the D-NC and B-NC datasets, the BIC defines the Gaussian and the Plackett copulas, respectively. These discrepancies are due to the number of parameters of the Student-*t* copula, and to the penalty that the BIC assigns to it. BIC penalty for the number of model parameters is greater than the one of AIC. Then, taking into account that the Student-*t* copula is a two-parameter copula, its penalization in the BIC is higher. In this way, as the adjustment of the Student-*t*, Gaussian, and Plackett copulas are quite similar for these two datasets, a greater penalty may lead to changes in the best-fit copula selection, as it happens in this case.

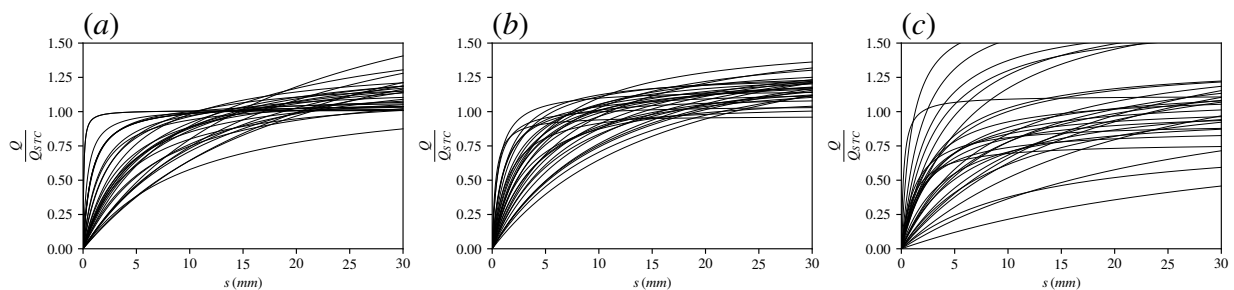
Third, observe from table 4-12 that, for the same dataset, some AIC values from different copulas are quite close, likewise some BIC values. Specifically, this is evident for the Student-*t* copula and Gaussian copula from the D-NC dataset, Student-*t* copula and Plackett copula from the B-NC dataset, or the Plackett copula and Frank Copula from the D-C dataset. These close values indicate that both copulas may be suitable for modeling the dataset since their adjustment is quite similar. As stated by Li et al. (2013), for the D-C dataset case where the Plackett copula is defined as the best-fit copula, Frank copula may be also a suitable copula for modeling the dependence between *a* and *b* parameters. In this sense, when the discrepancies between AIC/BIC values of two copulas are quite small, no major differences are expected in their behavior when modeling dependence between random variables.

To be in concordance with [Li et al. \(2013\)](#), the same best-fit copulas will be adopted in this work since this change does not have a major impact on the final curves.

Now, it is interesting to show a graphical comparison among the measured, simulated from the best-fit copula, and simulated without dependence, hyperbolic curves. Figure 4-11 shows the curves for the D-NC dataset, figure 4-12 for the B-NC dataset, figure 4-13 for the D-C dataset and figure 4-14 for the B-C dataset. Similar comparatives were performed by [Dithinde et al. \(2011\)](#); [Li et al. \(2012, 2013\)](#).



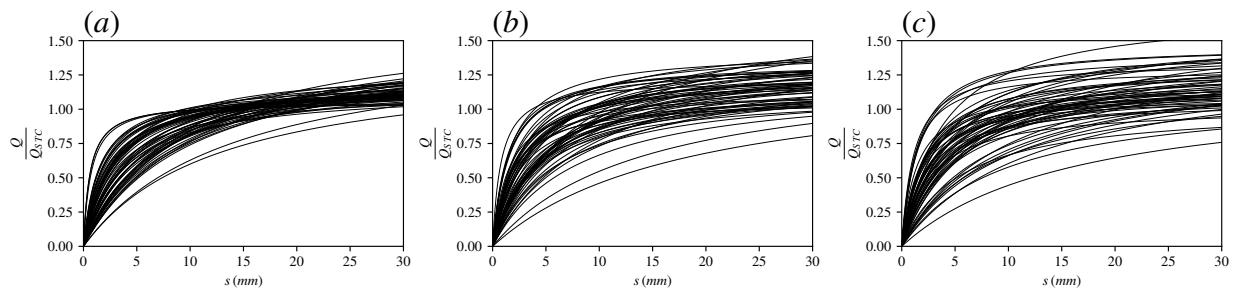
**Figure 4-11:** Hyperbolic curves of driven piles in noncohesive soils (D-NC). (a) measured, (b) simulated from a Gaussian copula, (c) simulated without dependence



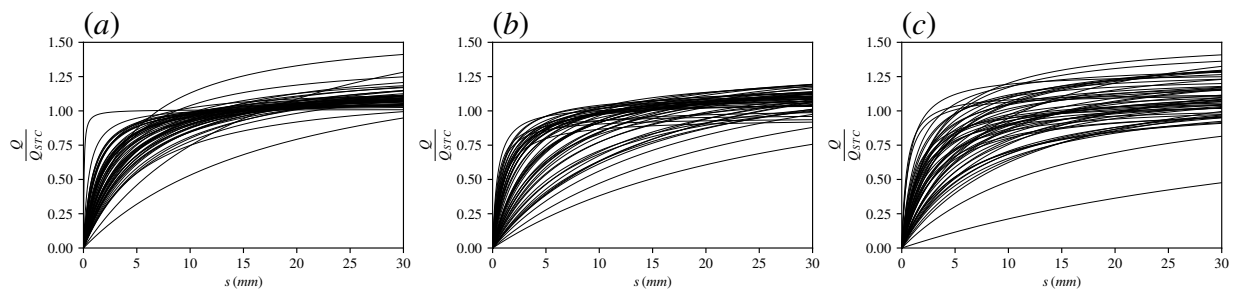
**Figure 4-12:** Hyperbolic curves of bored piles in noncohesive soils (B-NC). (a) measured, (b) simulated from a Plackett copula, (c) simulated without dependence

Similarities between measured and simulated with dependence hyperbolic curves are evident. Unlike assuming independence between  $a$  and  $b$  parameters of equation (4-34), copula theory can adequately represent their dependence and capture the remaining scatter of the normalized hyperbolic curves. In this way, copula theory demonstrates to be a suitable tool when capturing model uncertainty of normalized pile load-displacement curves. This applicability could be evaluated and extended to other types of geotechnical models where there may be dependence between the fitted parameters.

Nonetheless, it is worth mentioning that since the work conducted by [Dithinde et al. \(2011\)](#) the lognormal probability function has been adopted to represent both curve fitting parameters,  $a$  and



**Figure 4-13:** Hyperbolic curves of driven piles in cohesive soils (D-C). (a) measured, (b) simulated from a Plackett copula, (c) simulated without dependence



**Figure 4-14:** Hyperbolic curves of bored piles in cohesive soils (B-C). (a) measured, (b) simulated from a Plackett copula, (c) simulated without dependence

$b$ , without having evaluated other PDF. This assumption was replicated without validation by [Li et al. \(2012\)](#), and later by [Li et al. \(2013\)](#). Therefore, other univariate probability distributions have not been evaluated till the moment to model these two parameters. In this regard, [Li et al. \(2015a\)](#) demonstrated that other univariate probability distributions, different from the lognormal, such as weibull or gumbell truncated below zero distributions, are more suitable to represent the two hyperbolic curve fitting parameters of equation (4-32) or (4-33) in most cases. In this sense, by evaluating other marginal distributions, simulated curves of the figures 4-11, 4-12, 4-13, and 4-14, can get a better approximation to the measured data.

### 4.5.3 Geotechnical earthquake parameters

Earthquakes are one of the most destructive forces in nature and have a direct impact on civil structures and human life, so it is important to adequately characterize their properties in order to consider them in designs. From the geotechnical perspective, several ground motion parameters (GMP) have been documented, studied, and measured. Examples are peak ground acceleration (PGA), peak ground velocity (PGV), pseudo-spectral acceleration (PSA), cumulative absolute velocity (CAV), among others (see [Kramer et al., 1996](#), for more details).

Furthermore, it is common in practice to employ a ground motion prediction equation (GMPE) in order to predict peak ground motions parameters and response spectra of an earthquake, as functions of its magnitude, distance, and other properties. As noted by [Cornell \(1968\)](#) and [McGuire \(2004\)](#), GMPEs are critical elements when conducting probabilistic seismic hazard and risk analysis. Thus, quite a few GMPEs have been developed for different seismic regions and different GMPs, such as the PGA, PSA, CAV, etc.

In practice, it is common to independently predict the GMPs using GMPEs, and to assume that the GMPs are lognormally distributed ([Goda and Atkinson, 2009a](#)). That is to say, from the same seismic event, each GMP has its respective equation and the resulting parameters are assumed to follow a lognormal probability distribution. Nonetheless, a single GMP is not sufficient to correctly represent the behavior of ground motions, so multiple GMPs are needed to capture different features of an earthquake ([Xu et al., 2016b](#)). Additionally, given that the GMPs have a common source, attenuation and site attributes, these GMPs are indeed interrelated, and therefore, cannot be modeled as independent entities ([Goda and Atkinson, 2009a](#)).

In this sense, attempts have been made to characterize correlation among some GMPs. For this purpose, early works such as the one conducted by [Baker and Cornell \(2006\)](#) have employed the popular linear correlation coefficient. The application of Pearson's rho correlation coefficient was to be expected, given its simplicity and popularity, so several works have also employed it. For instance, the linear correlation among the spectral acceleration (SA), Arias intensity (AI), and PGA, was studied by [Baker et al. \(2007\)](#), while the linear correlation between the SA values at different periods of vibration was studied by [Baker and Jayaram \(2008\)](#), [Goda and Hong \(2008\)](#)

and [Goda and Atkinson \(2009b\)](#).

Disadvantages regarding linear correlation coefficients, and specifically Pearson's rho correlation coefficient, have already been exposed in this document. Furthermore, the multivariate normal distribution has been commonly employed for representing the dependence of multivariate ground-motion parameters ([Goda and Atkinson, 2009a](#)). From copula theory viewpoint, the multivariate normal model is not the only dependence structure available, and indeed it may not be the most adequate, so other dependence structures have to be assessed.

Although the importance of copula theory to model GMPs is evident, its use has been not so popular. One of the pioneering works was presented by [Goda and Atkinson \(2009a\)](#), where dependence between PGA and PSA parameters was studied for different vibration periods, employing copula theory. At the same time, [Goda and Atkinson \(2009a\)](#) validated the conventional approach, i.e., to fit lognormal distributions to the GMPs and to characterize their dependence through a linear correlation coefficient, by comparing it to the approach of copula theory. They employed a strong-ground-motion database, obtained from the Pacific Earthquake Engineering Research-Next Generation Attenuation of ground motions (PEER-NGA). Two elliptical copulas were initially considered, namely the Normal copula and the Student- $t$  copula. [Goda and Atkinson \(2009a\)](#) demonstrates through lognormal-probability-papers ([Ang and Tang, 2007](#)) and the AIC, that the lognormal probability distribution is suitable as the marginal distribution for both parameters, and also that the Normal copula is appropriate to represent their dependence. Furthermore, [Goda and Atkinson \(2009a\)](#) emphasizes in the benefits of using Kendall's tau coefficient to compute Pearson's rho through equation (4-18), instead of directly calculate Pearson's rho correlation coefficient. In this way, although the two-step approach was validated from the copula perspective, the benefits of using Kendall's tau coefficient are highlighted.

Another work in this framework was conducted by [Xu et al. \(2016b\)](#), in which the PGA and CAV ground motion parameters are studied and modeled as dependent parameters through copula theory. For this purpose, the Taiwan Strong Motion Instrumentation Program database was employed, and unlike [Goda and Atkinson \(2009a\)](#), several marginals distributions and copula functions were evaluated. However, [Xu et al. \(2016b\)](#) came to a similar conclusion to that of [Goda and Atkinson \(2009a\)](#), since it was found that the lognormal distribution is suitable as the marginal distribution for the GMPs and their dependence can be represented through a Gaussian copula. Thus, both studies validate the common two-step approach for their respective GMPs and database.

Additionally, [Xu et al. \(2016b\)](#) also constructed a copula model subject to a given magnitude and distance from a seismic event, using GMPEs for both PGA and CAV parameters. Normalized residuals between the predicted PGA and CAV parameters subject to a given magnitude and distance were also modeled through copula theory. Again, Gaussian copula and the lognormal distribution were also found suitable to model residuals of the PGA and CAV parameters, which is in concordance with the commonly used two-step approach.

More recently, [Cheng et al. \(2020\)](#) employed the vine copula theory to generate a multivariate

joint probability function of ground pseudo-spectral accelerations at different vibrations periods. The advantages of using vine copulas for the model were evident. Vine copulas not only better capture the multivariate dependence, but also manages to capture tail dependence between pairs of variables. The methodology proposed by [Cheng et al. \(2020\)](#) performs much better than the commonly employed multivariate normal distribution.

Finally, it is worth mentioning that the studies of ground-motion-parameters are characterized by having a substantial amount of data. It has already shown that the amount of data when studying shear strength parameters or settlement measures is commonly scarce; indeed, just on unusual occasions this amount of data exceeds a hundred. However, this is not the case of the GMPs, since in recent years the interest in earthquakes has grown by engineers, researchers, and especially, governments. For this reason, several programs around the world have been implemented to measure data related to GMPs in real-time, which allows having a large amount of information. One good example is the labor conducted by the *Pacific Earthquake Engineering Research Center*, where a large amount of geotechnical information (and of different kinds) on earthquakes has been compiled, analyzed and shared.

#### 4.5.4 Soil spatial modelling

Variability soils and rocks is evident and cannot be neglected. In fact, their properties may vary significantly, even for the same deposit. This variability is derived from different internal and external agents such as the deposition, post-deposition, weathering processes, anthropic actions, etc ([Zhu et al., 2017](#)). Thus, as it is impossible to exactly know the soil or rock properties in every single point of a deposit (or domain, mathematically speaking), and there will be intrinsic uncertainties associated with their characterization. So, it becomes mandatory to model soil or rock spatial variability and predict their properties to obtain a better description of their behavior under different circumstances.

Several methodologies are found in the literature for characterizing the spatial variability of materials, in this case soils or rocks. Nonetheless, not all these methodologies include the dependence among soil-rock parameters in their predictions, so they assume independence or in the best case a Gaussian dependence structure. Most recently, the applicability of copula theory for modeling spatial variability with dependent parameters has been evaluated by some pioneering works. So, the present section aims to review these particular methodologies and the role that copulas have performed.

Random field theory (see, e.g., [Vanmarcke, 2010](#)) constitutes one of the most comprehensive and popular methodologies applied for studying spatial fluctuation of material properties, such as the ones of soils and rocks. Succinctly, a random field is a function that defines random quantities of some variables over an arbitrary domain, which can be a multi-dimensional space. Thus, a random field is a function of the form  $H(x, \omega)$ , which associates a random quantity with a continuous index

$\mathbf{x} \in X$ , and a coordinate  $\omega$  in the sample space  $\Omega$ . Note that  $X \subset \mathbb{R}^d$ , for  $d \in \{1, 2, 3\}$ , is a bounded domain that represents the physical space.

Regarding dependence in the theory of random fields, two main types of correlations among variables are defined, namely the *cross-correlation* and the *auto-correlation*. The former describes the relationship between values of two (or more) random variables at the same point of the physical domain  $X$ . The latter describes the relationship among values of one random variable at two different locations in the domain  $X$ . Accordingly, observe that based on these definitions the dependence between shear strength parameters modeled in section 4.5.1 corresponds to a cross-correlation dependence.

Commonly, cross-correlation in geotechnical random fields is either obviated or taken into account through a cross-correlation matrix  $\mathbf{M}_c$ , where each term represent the correlation coefficient between the random variables. However, the importance of dependence has been elucidated in previous sections, so assuming independence may be misleading and lead to mistakes. Furthermore, cross-correlation matrixes  $\mathbf{M}_c$  have regularly implied the use of the Pearson's rho correlation coefficient and so, the multivariate Gaussian distribution. From copula theory perspective, to assume independence or Gaussian dependence, are just two of the many options of dependence structures available. In this regard, it is necessary to explore other dependence structures when modeling cross-correlated random fields. Two early studies have recently explored the use of copulas for this purpose.

After identifying these gaps, [Zhu et al. \(2017\)](#) proposed a new methodology based on copula theory to deal with the cross-correlation dependence among variables in geotechnical random fields. In the first instance, this early work generates independent multivariate anisotropic random fields through auto-correlation coefficients  $\rho_s$  obtained by a formulation of an elliptical scale of fluctuation. These auto-correlation coefficients serve to create an auto-correlation matrix  $\mathbf{M}_a$ , whereby some transformations, like the Cholesky decomposition, generate an independent random field.

Consequently, [Zhu et al. \(2017\)](#) proposed some procedures to transform the originally independent random field into a cross-correlated random field that follows the dependence structure of a given copula. Specifically, they incorporated the dependence structure of the Gaussian, Student- $t$ , Clayton, and Frank copulas into the cross-correlated random fields. The study performed a comparison between a copula-based cross-correlated random field of liquid limit, plasticity index and natural moisture content, and the original dataset of the same parameters. As a result of this comparison, [Zhu et al. \(2017\)](#) validated the proposed methodology since the statistics of both agree and, after the calibration, the copula-based cross-correlated random field generates a good representation of reality.

Finally, [Zhu et al. \(2017\)](#) also studied some incidences of the dependence parameter and type of selected copula over the final random fields. However, the scope this work is rather limited, so other studies complement it. [Wang et al. \(2020\)](#) further studied the incidences of copulas over the probabilities of failure of a homogeneous slope, under different conditions, modeled through the

copula-based cross-correlated random field proposed by [Zhu et al. \(2017\)](#). In this sense, [Wang et al. \(2020\)](#) employs the subset simulation algorithm ([Au and Beck, 2001](#)) to estimate the slope probabilities of failure by varying the copula selection, covariance of marginal distributions, the scale of fluctuation of the random field, and the cross-correlation coefficient.

Further details about the incidence of copulas in the final probability of failure of geotechnical structures are exposed in section 4.7. For the moment, what is truly important in this section is to highlight that cross-correlation plays a crucial role when modeling soil spatial properties. As noted by [Zhu et al. \(2017\)](#) and [Wang et al. \(2020\)](#), the independence assumption or a cross-correlation given by a Gaussian distribution leads to mistakes. So, it is necessary to evaluate different structures of dependence for modeling cross-correlated random fields, and copula theory shows to be suitable for this purpose.

Now, regarding the auto-correlation in the theory of geotechnical random fields, a new approach for its modeling based on copula theory was recently proposed by [Wang and Li \(2018\)](#). Note that auto-correlation is a different dependence concept from those studied so far in this document. Auto-correlation dependence is expected to be almost perfect for quite close points in the random field domain but simultaneously approaches zero for distant locations. In this sense, a copula parameter of dependence will be variable on the domain, and the selected copulas must be able to model from almost perfect dependences to almost null dependences.

The auto-correlation dependence structure of random fields has commonly been based on the assumption of gaussianity. However, this assumption is not always verified, and in the light of copula theory, Gaussian dependence is not the only option. Thus, [Wang and Li \(2018\)](#) evaluated the use of different dependence structures for a one dimensional random field, supported on copula theory. To this end they employed several copulas, including vine copulas.

[Wang and Li \(2018\)](#) emphasizes that the use of different dependence structures has a strong impact on the random field, except when this last one is highly ragged or highly smooth. Thus, a correct characterization and selection of the dependence structure (copula) is highly important when representing auto-correlation of a random field. Additionally, as the copula dependence parameters vary on the domain, candidate copulas to represent the auto-correlation dependence must be able to handle a wide range of dependence values, such as the comprehensive copulas ([Wang and Li, 2018](#)).

On the other hand, there are some methodologies worth mentioning related to the approaches of geospatial interpolation, such as the kriging prediction method. The majority of these approaches are based on some assumptions, such as a Gaussian dependence structure given by a variogram. However, this particular assumption has to be validated since it may not be an accurate representation of reality and leads to mistakes, as stated before.

Some pioneering works have explored the use of copulas in this regard. For example, [Marchant et al. \(2011\)](#) employed a Gaussian copula with generalized extreme value marginal distributions to

spatially model observations of soil cadmium concentrations, whose dataset exhibits outliers and tails. On the other hand, [Kazianka and Pilz \(2010\)](#) proposed three methodologies based on the kriging approach and copula theory to model a strongly skewed soil radioactivity measurements. They demonstrated that the proposed methodologies generalize some of those already existing in the state-of-the-art. Later, [Kazianka and Pilz \(2011\)](#) employed a Bayesian spatial interpolation method, in conjunction with copula theory, to model the same dataset employed in [Kazianka and Pilz \(2010\)](#), also exploring incidences of the uncertainties in fitting parameters.

There are other studies in the literature regarding the use of copulas in geospatial interpolation. An exposition of these works is out of the scope of this document, taking into account that their framework moves away from the geotechnical context. However, it is worth mentioning that a general observation of these studies is the transcendence of the dependence structure when interpolating geospatially. Outliers and skewness are common in geostatistics datasets, so their dependence structures are not correctly captured by the assumption of Gaussianity. Thus, it is necessary to evaluate other dependence structures, and copula theory is especially suitable for this end.

Finally, it is worth highlighting that the use of copulas in the modeling of soil spatial properties has not widely been studied, and further investigations must be carried out in this regard. Particularly, in the case of random fields, there are gaps in the state-of-the-art that can be explored through copula theory. For example, the auto-correlation in two or more dimensions considering copulas can be studied, or the implementation of the auto-correlation and cross-correlation in random fields, each one represented by copulas.

#### 4.5.5 Other uses of copulas in geotechnics

Copulas have been mostly employed in geotechnical engineering to model some types of dependence such as the one of shear strength parameters (section 4.5.1), settlement parameters (section 4.5.2), GMP (section 4.5.3) or spatial interpolation (section 4.5.4). However, there have been other applications of copulas in geotechnics that have not been as widely explored, and that will be mentioned in the following.

##### 4.5.5.1 Landslide hazard assessments

[Motamedi and Liang \(2014\)](#) evaluated the applicability of copulas for performing landslides hazard assessments in a regional scale. These assessments are characterized by three elements, namely magnitude  $M$ , frequency  $T$  and location (also called susceptibility)  $S$ . Assuming that these three elements are independent variables, which is the common practice, the following equation for hazard assessments is obtained ([Motamedi and Liang, 2014](#)):

$$H = P_m P_t S \quad (4-35)$$

where  $H$  is the landslide hazard,  $P_m$  is the probability of a landslide magnitude,  $P_t$  is the probability of a landslide recurrence, and  $S$  is the susceptibility or spatial probability.

Using a recorded landslide dataset at the Western Seattle area [Motamedi and Liang \(2014\)](#) validated instead the following equation:

$$H = P_t \cdot P(M \cap S)$$

in which  $T$  is modeled as an independent variable, but in turn  $M$  and  $S$  are modeled as dependent variables through the theory of copulas.

[Motamedi and Liang \(2014\)](#) evaluated a total of 14 families of copulas for the modelation of the dependence between  $M$  as  $S$ . They found that for this dataset the gumbell copula bests represents the structure of dependence between  $M$  and  $S$ . Furthermore, they performed a contrast between the results of the proposed methodology and the ones from equation (4-35), thus finding that the proposed approach makes a better representation of reality. Dependence, and hence copula theory, is suitable and even necessary in landslide hazard assessment on a regional scale ([Motamedi and Liang, 2014](#)).

#### 4.5.5.2 Earthquake early warning

Earthquake early warning (EEW) methodologies aim to detect earthquakes as soon as possible in order to disseminate alerts and prevent people before the peak ground motions arrival. For this purpose, the relationship between soil early motions (SEM) and peak ground motions (PGM) must be defined.

Nonetheless, there are uncertainties in this relationship, which lead to missed or false alarms. Missed alarms are defined as the event of SEM less than a prescribed threshold, but PGM greater than a triggering threshold. On the other hand, false alarms are defined as the event of SEM greater than a prescribed threshold, while the PGM is smaller than a triggering threshold.

Looking to model these uncertainties, [Wang et al. \(2018\)](#) employed a dataset of peak ground displacements within the first three seconds after  $P$ -waves arrives (as SEM parameter) and peak ground velocities (as PGM parameter), which are hereinafter referred to as  $PD_3$  and  $PGV$ , respectively. They evaluated six copula function to model the dependence between these two parameters, and found in the Frank copula its best representantion. Thus, this copula was calibrated from the dataset, and it was used to quantify the probability of missed and false alarms.

This copula model was compared with the conventional approach, which assumes independence between the parameters, and a better fit to the dataset and better results were found. In this way, taking into account the dependence between  $PD_3$  and  $PGV$ , [Wang et al. \(2018\)](#) showed that the probabilistic model of these two variables can be better characterized.

Additionally, [Wang et al. \(2018\)](#) proposed a copula-based methodology to optimize the selection of the triggering  $PD_3$ , thus reducing the probability of missed alarms ( $PGV$  greater than the threshold)

and false alarms (*PGV* less than the threshold). It is worth noting that, although it has not been yet done, this methodology could be extended to other types of geotechnical early warnings, such as the ones of debris flows or landslides.

#### 4.5.5.3 Studies on sample size

[Tang et al. \(2017\)](#) studied the impact of the sample size on the uncertainties in the definition of probabilistic geotechnical models. They assumed a true probabilistic model, which consists of a bivariate copula with two marginal distributions. From this model several samples of different size were drawn. For each sample size, a set of marginal distributions and copula functions are fitted to data and the best-fit marginal/copula functions were defined through the AIC methodology. Variation of the fitting parameters of each model are studied, and since the true model is known, it is possible to compare it with those models defined on each sample size.

As expected, the larger sample size the greater the probability of selecting the true copula functions, marginal distributions, and their parameters of adjustment. Furthermore, [Tang et al. \(2017\)](#) also found that the correlation coefficient has a great impact on the selection of the correct copula model. The stronger the dependence, the smaller the amount of data needed to correctly select true copula model, and vice versa, i.e. the weaker the dependence, the larger amount of data will be needed.

Furthermore, [Tang et al. \(2017\)](#) pointed out that another useful way to select the best copula and marginal distributions, for a dataset, is by comparing the variation of their AIC scores. As they pointed out, the best copula and best marginal distributions have less variation in their AIC scores. In order to obtain the AIC scores of the evaluated probabilistic models, and hence their variation, [Tang et al. \(2017\)](#) recommended the use of the Bootstrap method (see, e.g., [Efron, 1992](#)).

## 4.6 Uncertainties in the construction and definition of copula for geotechnical variables

It is quite unusual in geotechnical engineering to have a large amount of data to characterize soil and rock materials; thus, it is theoretically impossible to determine true probabilistic models. For this reason, in many cases, the statistics and models obtained from a sample, usually small, are assumed as obtained from populations. This assumption leads to uncertainties in modelling, which will be a function of the quantity and quality of information. These uncertainties have a direct impact on the veracity of the model and its results, such as the probability of failure.

Dealing with these uncertainties is not a trivial task, specially when the information is scarce. In any case, when it is not possible to reduce uncertainty, by collecting more and better information, a good practice is to characterize this uncertainty. Several investigations have been carried out

with the aim of characterizing uncertainty when constructing and selecting copula models that represent dependence among geotechnical random variables. An example of these methodologies is the work of [Tang et al. \(2017\)](#), who studied of the impact of sample size in the copula model definition.

In this section, we will study the work of [Li et al. \(2015a\)](#), who proposes an interesting methodology for characterizing model uncertainty in the construction and definition of multivariate probability models when the information is scarce. This methodology focuses on the characterization of uncertainty for the different probabilistic models employed by [Li et al. \(2013\)](#) for representing the pile load-displacement datasets of [Dithinde et al. \(2011\)](#). Recall that the works of [Dithinde et al. \(2011\)](#) and [Li et al. \(2013\)](#) were already introduced in section 4.5.2. Furthermore, it is worth mentioning that the procedures and remarks of this section can be extended to other examples of dependence of geotechnical random variables when the information is scarce.

For this purpose, the bootstrap method ([Efron, 1992](#)) was employed by [Li et al. \(2015a\)](#) for studying uncertainties in both copulas and marginal distributions. Bootstrap method belongs to the so-called resampling techniques, which are useful for deriving sampling distributions of statistics from the measured data.

Briefly, the idea behind the bootstrap method consists of creating  $N_s$  sets of bootstrap samples with the same sample size  $N$  of the original dataset. Each bootstrap sample is constructed by random sampling with replacements from the original dataset; that is to say, every single observation can appear once, more than once, or not at all. For each bootstrap sample, estimates such as the mean, standard deviation, copula parameter of dependence, and even AIC/BIC values, are computed. After obtaining all values of the desired measure, every estimative can be associated with a sampling distribution, which represents its uncertainty. Readers are referred to [Most and Knabe \(2010\)](#) and [Luo et al. \(2013\)](#) for additional applications of the bootstrap method in geotechnical engineering.

According to [Li et al. \(2015a\)](#), there are two kinds of uncertainty when constructing copula models: one related to the adjustment of the parameters of the marginal distributions and copula functions, and another related to the definition of the best-fit marginal distribution and copula function. [Li et al. \(2015a\)](#) employed the bootstrap method to characterize the uncertainty of the sample statistics of the marginal distributions for different pile load-displacement databases, i.e. mean and standard deviation. Furthermore, this work also characterizes, through the bootstrap method, the uncertainty in the AIC scores when defining the best-fit models for both marginal distributions and copula functions.

Uncertainty in the copula dependence parameter was mentioned, but not explicitly studied, by [Li et al. \(2015a\)](#). Therefore, this uncertainty will be explored in this document for illustration of the bootstrap procedure and in order to complement the original work of [Li et al. \(2015a\)](#). In this way, the database of [Dithinde et al. \(2011\)](#) is adopted for adjusting a set of copula functions and to conduct a bootstrap procedure in order to characterize the uncertainty in the definition of the fitting parameters of these copulas. The copulas to be used in this procedure will be the Gaussian,

Plackett, Frank, and Nelsen No. 16.

Note that the Student- $t$  copula is not included in the analysis since, by employing the method of moments for the adjustment, its parameter of dependence is the same as the Gaussian copula. Therefore, results obtained for the Student- $t$  copula would be the same of those of the Gaussian copula.

As recommended by [Most and Knabe \(2010\)](#) and [Luo et al. \(2013\)](#),  $N_s = 10000$  is adopted in the bootstrap procedure conducted for the evaluation of the uncertainties in the dependence parameter of each copula. For each bootstrap sample, copula dependence parameters are computed by the method of moments, i.e., employing respective Kendall's tau coefficient and equation (4-19) or the corresponding simplifications exposed in section 4.4.4. So, at the end of the procedure there will be 10000 estimations of the dependence parameters to construct every sampling distribution.

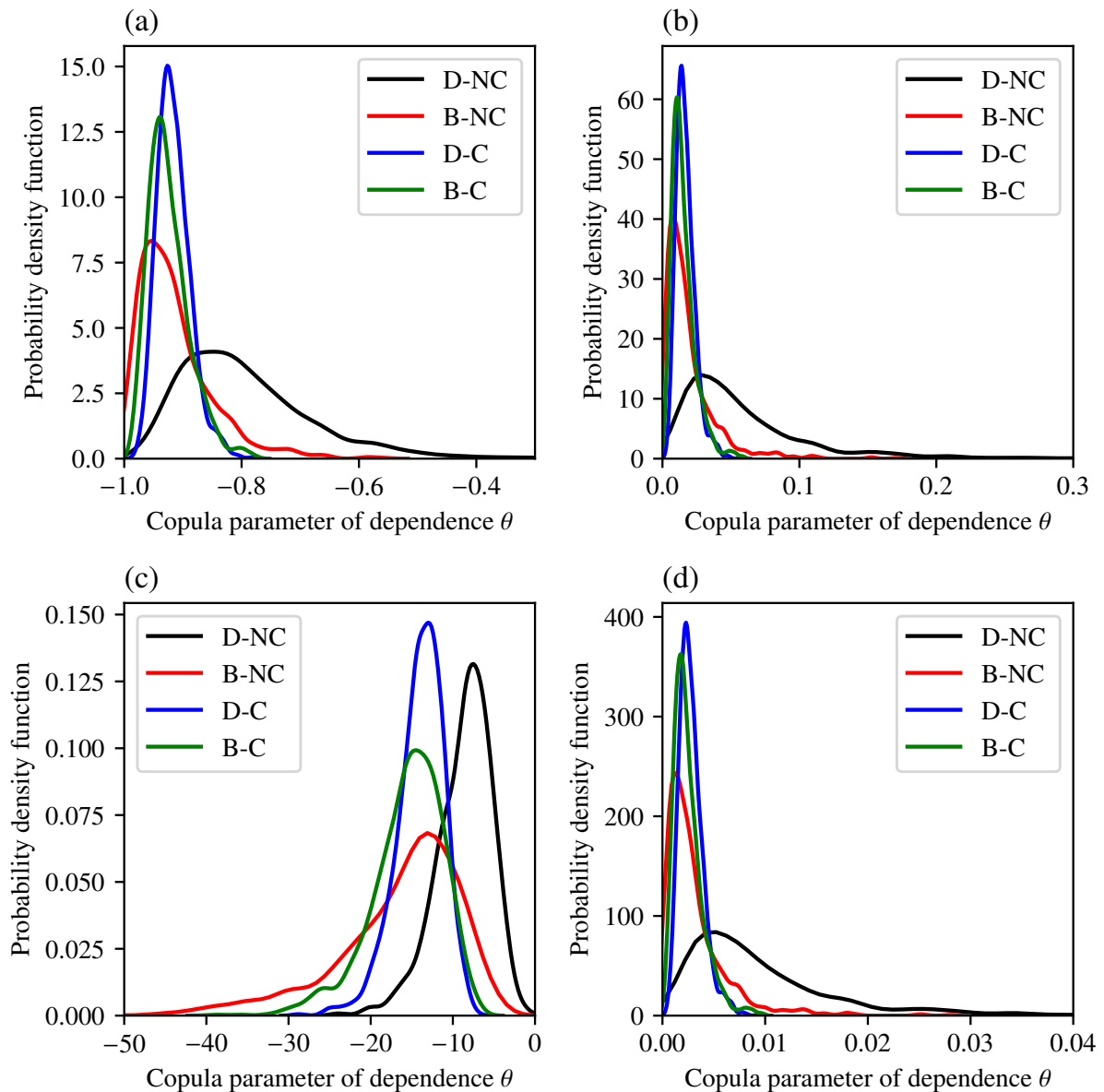
Resulting sampling distributions for the parameters of dependence of this example can be observed in figure 4-15, where each graph corresponds to a copula, and each curve to a pile test category. Statistics from the estimated distributions are computed and summarized in table 4-13.

**Table 4-13:** Mean and coefficient of variation C.O.V. from the obtained sampling distributions for the copula dependence parameters in the different pile tests of [Dithinde et al. \(2011\)](#).

Pile Test	N	Gaussian		Plackett		Frank		No 16	
		Mean	C.O.V.	Mean	C.O.V.	Mean	C.O.V.	Mean	C.O.V.
D-NC	28	-0.795	0.135	0.068	0.797	-8.643	0.395	0.010	0.842
B-NC	30	-0.914	0.068	0.029	0.961	-16.720	0.519	0.003	0.992
D-C	59	-0.916	0.030	0.017	0.418	-13.964	0.200	0.003	0.402
B-C	53	-0.923	0.040	0.020	0.583	-15.408	0.291	0.003	0.600

From figure 4-15 and table 4-13 some important remarks can be extracted. Firstly, by comparing tables 4-11 and 4-13, it is concluded that the bootstrap mean estimates of copula parameters of dependence are quite similar to those obtained from the original dataset. This fact indicates that distributions obtained through the bootstrap method can appropriately capture the characteristics of the data set.

Additionally, it is evident that the larger the dataset, the lower the dispersion of the models. This fact is noticeable by comparing, in figure 4-15, the curves of the D-NC data set (28 samples), with the ones of the D-C dataset (59 samples). Furthermore, these comparisons are quantified in table 4-13 where the coefficient of variation C.O.V. of the sampling distributions is higher when the dataset is smaller. Thus, the highest C.O.V. corresponds to the D-NC dataset, and the B-NC, B-C, and D-C data sets follow in order, which coincides with the amount of information of each pile load-displacement test. Therefore, this fact constitutes a motivation to collect more information, in order to obtain more accurate results.

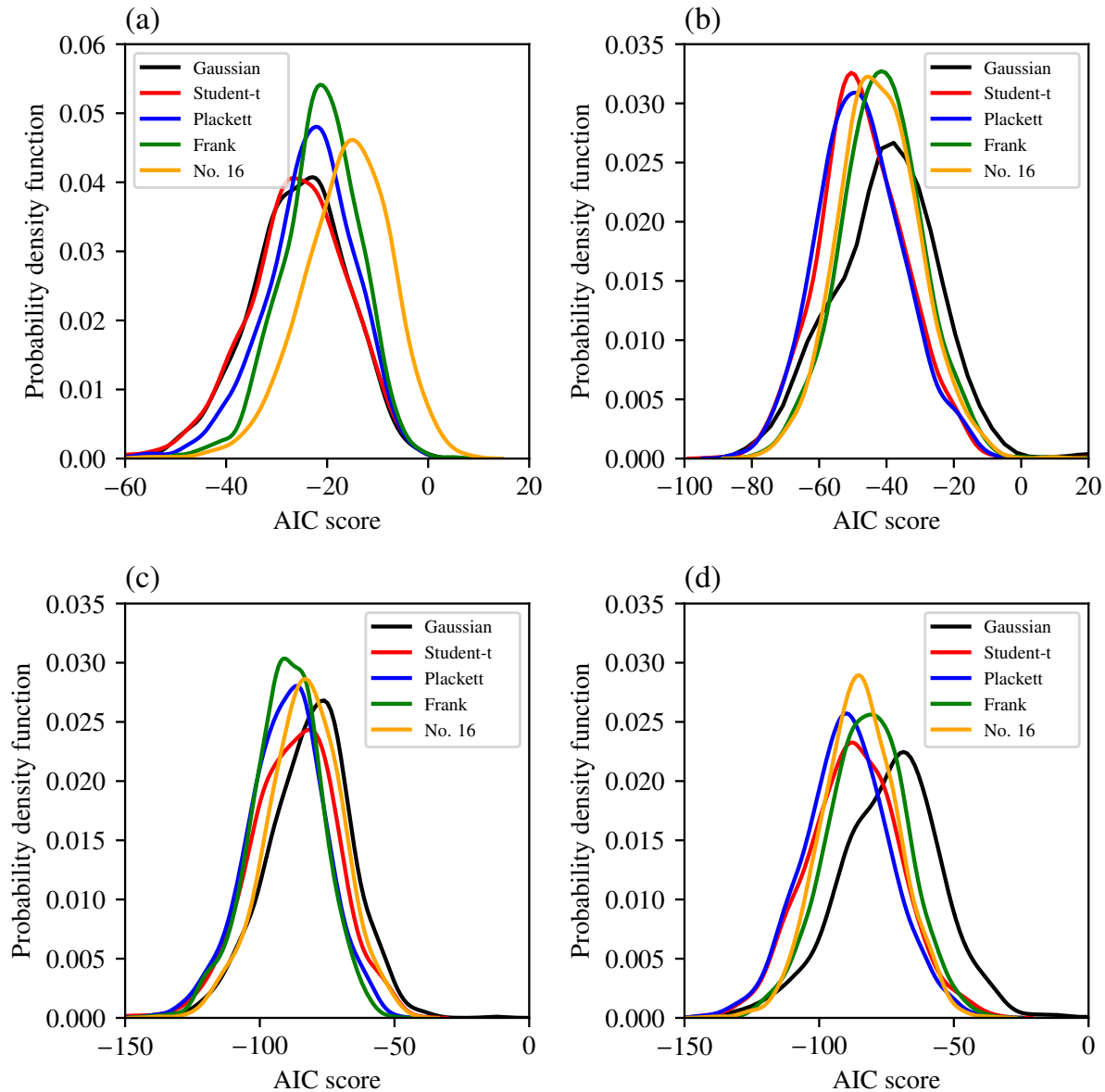


**Figure 4-15:** Sampling distributions for the parameters of dependence of different copulas and for the different pile tests of [Dithinde et al. \(2011\)](#). (a) Gaussian copula, (b) Plackett copula, (c) Frank copula, and (d) Nelsen No 16 Copula

On the other hand, uncertainty when selecting the best-fit marginal/copula models was also studied by [Li et al. \(2015a\)](#) through the computation of the AIC values for each copula in each pile dataset. It is interesting summarize some of these results here. Again, using  $N_s = 10000$ , a bootstrap procedure over the dataset of [Dithinde et al. \(2011\)](#) is conducted. For each bootstrap sample, AIC scores for each copula are computed and the corresponding best-fit copula is defined. So, there will be 10000 AIC scores for each copula and 10000 best-fit copulas for each load-displacement

test at the end of the procedure.

Figure 4-16 presents the sampling distribution of the AIC values for different copulas and pile tests. Additionally, table 4-14 contains the number of times that each copula is identified as the best-fit copula for each pile test.



**Figure 4-16:** AIC probability density functions of different copulas for different piles. (a) D-NC dataset, (b) B-NC dataset, (c) D-C dataset, and (d) B-C dataset

Figure 4-16 demonstrates that the bootstrap mean estimates of the AIC scores match well with those values obtained from the original dataset. However, AIC bootstrap sampling distributions

**Table 4-14:** Number of times that each copula is defined as the best fit copula for each pile test category

Pile test	Gaussian	Student- <i>t</i>	Plackett	Frank	No 16
D-NC	<b>5511</b>	3610	750	48	81
B-NC	1600	3216	<b>4204</b>	364	616
D-C	861	1470	3327	<b>4207</b>	135
B-C	390	2188	<b>5589</b>	222	1611

for each copula are broad, which means that there may be large uncertainties when identifying the best-fit copula to represent the dependence between  $a$  and  $b$  parameters. Again, uncertainties (which are related to coefficients of variation C.O.V. values) are a function of the amount of data. Thus, the more available information the smaller coefficient of variation., and therefore the lower uncertainty.

Table 4-14 shows that no copula can be identified as the best-fit copula with total certainty, that is to say, with a probability of one or almost one. There are copulas with a higher probability of being the best-fit copula, such as the Gaussian in the D-NC pile test, the Plackett in the B-NC and B-C pile tests, or the Frank in the D-C pile test. However, there is no total certainty in the copula selection, and therefore this uncertainty must be considered in the models.

Additionally, note from table 4-14 that copulas with a major probability of being the best-fit copula mostly agree with those defined in table 4-12. However, interestingly this is not the case of the D-C pile dataset, where Frank copula has the major probability of being the best-fit copula. Li et al. (2013) pointed out that the Frank copula, instead of the Plackett copula, may also be suitable to model the D-C dataset, and this fact is verified by this bootstrap procedure of the present section. Thus, the bootstrap method also demonstrates to be quite useful to define best-fit copula by taking into account model selection uncertainties.

Li et al. (2015a) emphasizes the importance of the bootstrap method to perform reliable judgments when selecting best-fit copulas, especially when the sample size is small, and therefore the uncertainties large. Uncertainties must be incorporated on the selection of the best-fit copulas and the estimation of model parameters. Hence, the bootstrap method proves to be an adequate tool.

Finally, it is worth mentioning again that although the work conducted by Li et al. (2015a) is applied to the database collected by Dithinde et al. (2011), it could also be extended to other geotechnical parameters. Concepts of model uncertainties are quite general, and so also the results presented in this section.

## 4.7 Reliability analysis in geotechnics considering dependence through copulas

Copulas have been employed in the reliability analysis of different geotechnical structures and models. Some examples are infinite slopes (Wu, 2015; Zhang et al., 2014), slopes with a circular failure mechanism (Xu et al., 2016a; Wu, 2013a), foundations (Wu, 2015, 2013b), rock wedge slopes (Li et al., 2015b), retaining walls (Tang et al., 2013; Phoon and Ching, 2014; Li et al., 2015b), among others. However, even though all these studies employ copulas, and indeed have several similarities, they have different methodologies and scopes that nurture the state-of-the-art.

It is necessary to explore and expose these works and their implications on the state-of-the-art about the use of copulas in reliability analysis in geotechnical engineering. Thus, this chapter presents the most crucial remarks and findings in this framework so far. Some examples and graphs will be replicated to highlight some cornerstones in the reliability analysis using copulas in geotechnical engineering.

Let us recall that a probability of failure can be understood as the probability of exceeding a given threshold or limit state, which leads to unwanted behaviors or instability. Mathematically speaking, the probability of failure is defined as the probability measure of a failure domain, which is delimited by a limit state function  $G(X_1, \dots, X_d)$ . Some examples of limit state functions are:  $G = FS - 1.0$ , where  $FS$  is the factor of safety, or  $G = R - S$ , where  $R$  are the resistance forces and  $S$  the applied forces.

In a general case, the probability of failure can be expressed through the following equation:

$$P_f = \int \cdots \int_{G(X_1, \dots, X_d) \leq 0} f(x_1, \dots, x_d) dx_1 \dots dx_d \quad (4-36)$$

where  $f(\cdot)$  is the probability density function of the random variables  $x_1, \dots, x_d$ , and  $G \leq 0$  is the failure region.

In a bivariate case, and considering that the random variables of the model are the shear strength parameters of the Mohr-Coulomb failure criterion, i.e., cohesion and friction angle, the probability of failure ( $P_f$ ) of the geotechnical model can be expressed as:

$$P_f = \iint_{G(c, \phi) \leq 0} f(c, \phi) dc d\phi \quad (4-37)$$

where  $f(c, \phi)$  is the bivariate probability density function of cohesion and friction angle, and  $G \leq 0$  represents the failure domain.

### 4.7.1 The copula based sampling and copula direct integration methods.

In order to conduct reliability analysis, considering dependence through copula theory, two major procedures have to be conducted. First, it is necessary to define the random variables involved in the geotechnical system, evaluate their dependence, and construct the multivariate probability distribution using copula functions. Details of these procedures are shown in section 4.4, and examples are presented in section 4.5. Second, a methodology must be employed to compute the probability of failure of the geotechnical model in order to assess its reliability.

There exist several methodologies in the literature to address this second step. Some of them are direct integration, transformations, or simulations. Each one of these alternatives has its particularities, advantages, and disadvantages. For instance, direct integration is accurate but may be unfeasible in very high dimensions or in very complex models, meanwhile simulation can be employed for complex models but may be too computationally expensive when the probability of failure is low, and inaccurate when the number of simulations is not enough.

In the state of the art, two main techniques are used to estimate the probability of failure of geotechnical models when dependence through copula theory is considered: the direct integration method and the Monte-Carlo simulation. It is worth mentioning that these techniques have been renamed by some works as the *copula direct integration method (CDIM)* (see, e.g., Tang et al., 2013, 2015; Xu et al., 2016a) and the *copula based sampling method (CBSM)* (see, e.g., Wu, 2013a,b; Xu and Zhou, 2018), respectively, this considering that the direct integration and the Monte-Carlo simulation are applied on a PDF defined through copulas. Nonetheless, in this document we will continue using the conventional names of these methodologies, since the CDIM and the CBSM may be misleading.

#### 4.7.1.1 Direct integration method in copulas

Let us recall that the probability of failure  $P_f$  can be computed through the Riemann–Stieltjes integral of equation (2-4). When the random variables of the model are just soil cohesion and friction angle, and its structure of dependence is given by a copula  $C$ , equation (2-4) can be expressed as

$$P_f = \iint_{[0,1]^2} \mathbb{I}[G(c, \phi) \leq 0] dC(c, \phi; \theta) \quad (4-38)$$

where  $G$  is the limit state function of the model.

If  $C$  is sufficiently differentiable and the copula density function  $c$  exists, then equation (4-3) can be replaced in equation (4-38), and hence, the latter can be rewritten as:

$$P_f = \iint_{G \leq 0} f_c(c) f_\phi(\phi) c(F_c(c), F_\phi(\phi); \theta) dc d\phi \quad (4-39)$$

Thus, in this case the direct integration method aims to directly integrate equation (4-39) in order to obtain the probability of failure of the given geotechnical model. Until this moment, the direct integration method has only been applied over bivariate copulas constructed to represent dependence of the shear strength parameters (see, e.g., [Tang et al., 2013](#); [Phoon and Ching, 2014](#); [Tang et al., 2015](#); [Xu et al., 2016a](#)). However, its applicability could also include other geotechnical parameters, such as the unit weight, or other constitutive models.

This methodology was initially applied in geotechnics by [Tang et al. \(2013\)](#) for the reliability analysis of an infinite slope and a retaining wall. Subsequently, other works such as the ones conducted by [Phoon and Ching \(2014\)](#), [Tang et al. \(2015\)](#) or [Xu et al. \(2016a\)](#) have employed it, also in a geotechnical context. Its concept is quite straightforward and can be applied in many geotechnical models. However, the direct integration method has some limitations related to the complexity of the model or the number of random variables.

Note that the double integral of equation (4-39) can become too complex, reason why the first derivative of the copula function is employed by some researchers in order to simplify its computation, that is to say:

$$M(u_1, u_2; \theta) = \frac{\partial C(u_1, u_2; \theta)}{\partial u_2} \quad (4-40)$$

So, by replacing equation (4-40) in equation (4-39), this latter can be rewritten as (see more details at [Tang et al., 2013](#)):

$$P_f = \int_{FS \leq 1} f_\phi(\phi) M(F_c(c(\phi)), F_\phi(\phi); \theta) d\phi$$

Note that for further derivation, the expression of  $c$  in terms of  $\phi$  must be known, or vice versa, according to the case and with which variable is being integrated. This expression must be derived from the performance function, and it is here where some complications may appear. In some cases, this derivation is straightforward, such as in an infinite slope model, but in some others, this expression is untractable and some approximations must be applied.

An example of these approximations is the *g-line*, proposed by [Klar et al. \(2011\)](#). The *g-line* is basically a simplified mathematical approximation to a complex performance function  $G$  whose analytical expression is unknown or untractable. By using the *g-line*, the complexity of the model can be significantly reduced and, therefore, the computation of the probability of failure simplified.

In geotechnical engineering, [Xu et al. \(2016a\)](#) employed the *g-line* in order to represent the limit state function of a homogeneous slope with a circular failure surface. [Xu et al. \(2016a\)](#) found in a quadratic polynomial equation a good approximation to the *g-line*, so the expression of  $\phi$  in terms of  $c$  was derived from it. In other example, the Shengjin's formulas ([Fan, 1989](#)) were employed by [Phoon and Ching \(2014\)](#) and [Tang et al. \(2013\)](#) to solve the cubic equations of a retaining wall model and to express  $c$  as a function of  $\phi$  in order to conduct the direct integration.

In this way, equation (4-39) has the advantage of computing the probability of failure in an accurate manner, but its applicability is not that wide. There are no errors regarding the linearization at the design point, such as in FORM and SORM, or statistical errors regarding small sample sizes such as in the MCS, which are aggravated in small probabilities of failure calculations (see [Tang et al., 2013](#); [Phoon and Ching, 2014](#)). However, some complications emerge when the performance function is strongly non-linear, or when it depends on more than two variables, which increases the complexity of the equation (4-37), since its dimension increases.

From our point of view, the approximation to the  $g$ -line applied by [Xu et al. \(2016a\)](#) is a good and practical alternative for complicated geotechnical models where the direct integration method is wanted to be applied. Nonetheless, this approximation requires hundreds of runs of the geotechnical system to search the critical state (e.g., where  $FS = 1$  or  $R = S$ ). On the other hand, the coefficient of determination  $R^2$  of the fitted  $g$ -line equation is generally far from perfect, leading to a loss of accuracy in the final estimates of  $P_f$ . Finally, this methodology is difficult to apply in larger dimensions. Consequently, the direct integration method requires more investigation to become a more general methodology, that can be applied in more complicated geotechnical models.

#### 4.7.1.2 Monte Carlo Simulation in copulas

In our context, Monte Carlo Simulation MCS aims to approximate equation (4-36) or (4-37) through a repeated random sampling. Specifically, MCS draws samples from the copula that represents the dependence structure of the model, after that, these samples are transformed to the space of the basic variables through some method for estimating the inverse of the CDF of the marginal distributions. Then, the performance function  $G$  is evaluated for each one of these samples, and finally the probability of failure is computed as the ratio between the number of samples that leads to failure over the total number of samples.

Some researches, have highlighted the statistical error due to the number of samples underlying MCS in geotechnical reliability analyses (e.g., [Tang et al., 2013](#); [Phoon and Ching, 2014](#)). This statistical error is aggravated in too small probabilities of failure since more samples are needed for accurate calculations. In this regard, it is true that the lower the probability of failure, the larger the sample size must be to avoid inaccuracies, and this may be time consuming.

This issue has been subject to an extensive discussion by different researchers since the number of samples of a MCS considerably affects the accuracy of the reliability analyses. In this regard, there exists several methodologies and equations to estimate the minimum sample size in a MCS. For example, equation (2-8) is one of these estimates, and it is particularly one of the most straightforward.

It is beyond the scope of this document to further discuss the minimum sample size of a MCS and the repercussions of insufficient samples. So, readers interested in this topic are referred to [Ang and Tang \(1984\)](#), and the references therein cited, for a more comprehensive introduction and details.

Additionally, it is worth mentioning that there are currently other methodologies of simulation especially useful for too small probabilities of failure, such as subset simulation (see [Au and Beck, 2001](#)) or line sampling ([Schuëller et al., 2004](#)). Some pioneering works have already integrated some of these methods of simulation on the reliability analyses of geotechnical structures that employ copulas (see e.g. [Wang et al., 2020](#)).

Despite this fact, the MCS constitutes a straightforward method with many advantages, even over the direct integration method exposed by [Tang et al. \(2013\)](#). The MCS is perhaps one of the best current alternatives to conduct reliability analysis in the framework of geotechnical engineering when considering dependence through copula theory. The MCS does not need a linearization at the design point such as the FORM and SORM, neither an approximation to the limit state function for complex geotechnical problems such as the direct integration. Additionally, it can handle systems with many random variables, i.e., geotechnical systems of high dimensions.

From our point of view, the MCS is currently a very good alternative to conduct reliability analyses of geotechnical models where the dependence is considered. However, this methodology may be computationally expensive, especially for high dimensions and fairly small probabilities of failure. More advanced methodologies, such as subset simulation (see [Au and Beck, 2001](#)) or line sampling ([Melchers and Beck, 2018](#)), exists. [Alvarez et al. \(2018\)](#) presents a framework that allows to use subset simulation with copulas and random set theory, in order to conduct efficient reliability analysis where dependence and both aleatory and epistemic uncertainties are considered. Particularly in geotechnical engineering, [Wang et al. \(2020\)](#) employed copulas in conjunction with subset simulation, in order to conduct reliability analysis of geotechnical structures. Furthermore, chapter 6 is dedicated to present a methodology in which subset simulation can be employed in conjunction with copula theory.

### 4.7.2 Results of reliability analysis

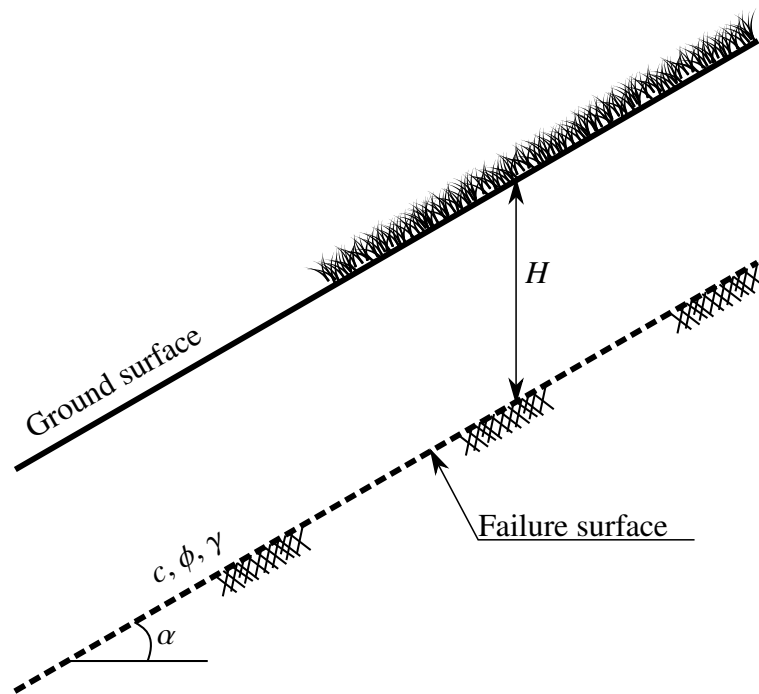
Copulas have been used to estimate the probability of failure of several geotechnical structures and systems, for instance, infinite slopes, foundations, retaining walls, among others. So, it is important to summarize some of these results and remarks, which help us to understand how copulas impact reliability analysis of geotechnical models.

For this purpose, the infinite slope example studied by [Tang et al. \(2013\)](#) is replicated in this document. This model is selected given its simplicity and that the incidence of copulas in the probability of failure has already been studied by [Tang et al. \(2013\)](#) and [Tang et al. \(2015\)](#) under different conditions.

Similar outcomes and conclusions, from this infinite slope example, are obtained from other geotechnical structures, such as retaining walls ([Phoon and Ching, 2014](#); [Tang et al., 2013](#)), slopes with a circular failure mechanism ([Wu, 2013a](#); [Xu et al., 2016a](#)) and three dimensional slopes ([Xu and Zhou, 2018](#)). Also, similar conclusions appear when there are more than two dependent ran-

dom variables (Wu, 2013b), or when evaluating not just one failure mode but the failure of the entire geotechnical system (Li et al., 2015b).

Figure 4-17 presents a scheme of the infinite slope model to be analyzed. Here,  $H$  stands for the thickness of the unstable layer,  $\gamma$  represents the soil unit weight,  $\alpha$  is the slope angle relative to the horizontal plane, and  $c$  and  $\phi$  are soil cohesion and friction angle, respectively. Note that the groundwater table is assumed to be under the failure surface, and is not considered in the present analysis. Additionally, neither overload nor seismic loads are considered in this slope model.



**Figure 4-17:** Infinite slope model

The factor of safety of the infinite homogeneous slope shown in figure 4-17 is given by

$$FS = \frac{c + \gamma H \cos^2 \alpha \tan \phi}{\gamma H \sin \alpha \cos \alpha} \quad (4-41)$$

The performance function for carrying out the reliability analysis will be of the form  $G = FS - 1$ , where the  $FS$  is the factor of safety obtained from equation (4-41).

In equation (4-41), both  $c$  and  $\phi$  will be taken as random variables, and their dependence will be modelled through a copula. The other parameters of equation (4-41) are adopted as deterministic.

We will do three modifications to the original work of Tang et al. (2013) related to the copula dependence parameter, the procedure to compute the probability of failure, and the evaluation of

an additional copula. Pearson's rho correlation coefficient was employed by Tang et al. (2013) and Tang et al. (2015) to estimate the dependence parameter of the different copulas used in the reliability analysis. However, in this document Kendall's tau coefficient is employed, given that it is a more suitable measure to represent the dependence in copula theory (see section 4.4.2). On the other hand, the direct integration was employed in the original example to compute the probability of failure. Nonetheless, the MCS will be employed for this purpose in this document, since it is a more general methodology, and it is also interesting to explore its performance and compare it with direct integration. Finally, in addition to the four copulas proposed by Tang et al. (2013), the Student- $t$  copula is evaluated, following the recommendations of Embrechts et al. (2001). In this way, the Gaussian, Student- $t$ , Plackett, Frank, and Nelsen No. 16 copulas, will be employed in this example.

Regarding the other aspects from the original example of Tang et al. (2013), these will be addressed without modifications. Thus, the lognormal probability distribution is adopted to model both  $c$  and  $\phi$ , where the mean and coefficient of variation C.O.V. of  $c$  are taken as 12 kPa and 0.4 respectively, and the mean and coefficient of variation C.O.V. of  $\phi$  are taken as  $30^\circ$  and 0.2 respectively. On the other hand, the deterministic parameters are taken as  $H = 5$  m,  $\alpha = 30^\circ$  and  $\gamma = 17$  kN/m<sup>3</sup>. Finally, the Kendall's tau correlation coefficient  $\tau$  between  $c$  and  $\phi$  is  $-0.5$ .

Following the procedure of Tang et al. (2013), the incidence of copulas in the probability of failure is studied under three conditions, namely: (1) geometry ( $H$  and  $\alpha$ ), (2) C.O.V. scaling factor of the marginal probability distributions ( $C.O.V./\lambda_{c,\phi}$ ) and (3) correlation between  $c$  and  $\phi$  ( $\tau_{c,\phi}$ ).

Changes in each condition will be represented as changes in the *nominal factor of safety*  $FS_n$  (Tang et al., 2013). The  $FS_n$  are computed following the algorithm presented by Tang et al. (2013) or Phoon and Ching (2014). In this algorithm, Plackett copula is employed to simulate random values of the shear strength parameters, which in turn serve to compute as many  $FS$  as random samples. Each  $FS_n$  is taken as the 5% fractile of all the obtained  $FS$  for the fixed condition. More details about  $FS_n$ , and the algorithm to compute it, can be found in the aforementioned two references.

The nominal factor of safety has the advantage of taking into account the coefficients of variation of the distributions of the shear strength parameters, and also the correlation between these two variables. This fact contrasts with a simple factor of safety, where only the mean values of the random variables are employed, so there is no room for uncertainty in the estimates of the factor of safety.

Unlike Tang et al. (2013), in this document nominal factors of safety will vary between 1.0 and 1.16. That is to say, when plotting  $FS_n$  against the probability of failure of each copula, the horizontal axes ( $FS_n$ ) for all the conditions are the same (i.e.,  $FS_n \in [1.0, 1.16]$ ), and their values are due variations of each particular condition. Specifically,  $H \in [3.84\text{m}, 6.27\text{m}]$ ,  $\alpha \in [27.7^\circ, 32^\circ]$ ,  $\lambda_{c,\phi} \in [0.67, 1.64]$  and  $\tau_{c,\phi} \in [-0.78, -0.26]$ , and all these variations will lead to the same horizontal axis of  $FS_n$ .

Additional to the four copulas initially evaluated by Tang et al. (2013), namely Gaussian, Plackett, Frank and Nelsen No. 16 copulas, Student- $t$  copula is also studied in this document. Note that the aforementioned copulas are able to be adjusted to the Kendall's tau coefficient assumed in this example.

Then, MCS is performed for each copula, thus defining the probabilities of failure for aforementioned conditions. A total number of  $10^7$  samples will be simulated and employed by each MCS. The above number is not arbitrary since, from Tang et al. (2013), the orders of magnitude of the probabilities of failure of the infinite slope example are known. So, by applying equation (2-8), a minimum sample size of  $10^7$  is obtained for the smallest probabilities of failure.

After applying the above procedure with all the studied copulas the probabilities of failure are obtained and presented in figure 4-18. In this figure, each plot corresponds to the variation of a particular parameter, and each curve corresponds to a copula.

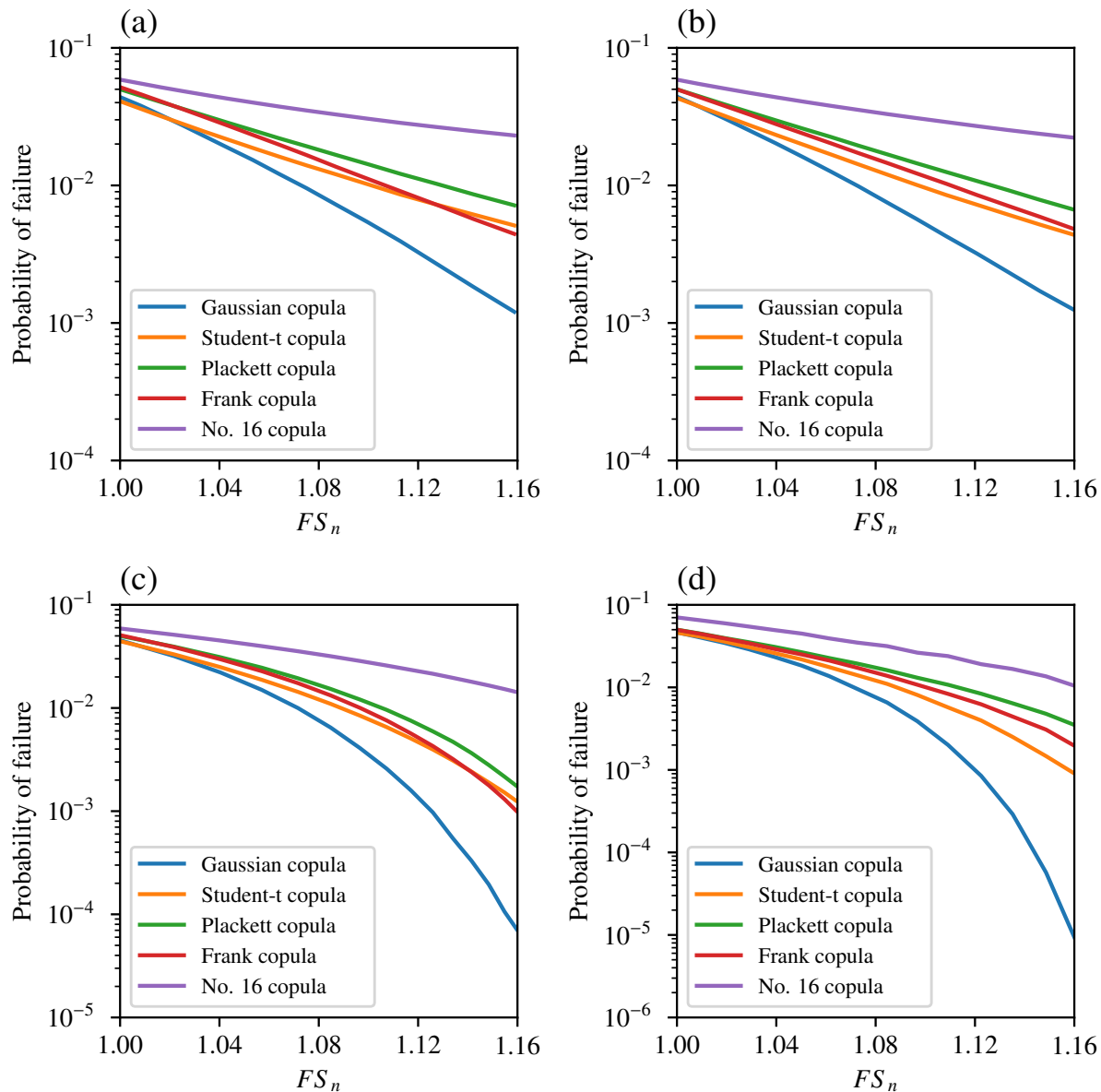
Additionally, table 4-15 presents a comparison of the probabilities of failure obtained through all the copulas, for all the studied failure mechanisms. For constructing table 4-15, probabilities of failure from the Gaussian copula are taken as reference, and the ratio  $P_{f_i}/P_{f_{\text{gaussian}}}$  is computed, in which  $P_{f_i}$  is the probability of failure obtained from each respective copula. Thus, all computed ratios are summarized in table 4-15.

**Table 4-15:** Comparison of the  $P_f$  obtained from different copulas for the infinite slope example, and taking the  $P_f$  of the Gaussian copula as reference for computing the ratio  $P_{f_i}/P_{f_{\text{gaussian}}}$ .

Copula	$H = [6.27, 3.84]$			$\alpha = [32, 27.7]$			$\lambda = [0.67, 1.64]$			$\tau = [-0.26, -0.78]$		
	1.00	1.08	1.16	1.00	1.08	1.16	1.00	1.08	1.16	1.00	1.08	1.16
Gaussian	1	1	1	1	1	1	1	1	1	1	1	1
Student- $t$	0.94	1.69	4.35	0.98	1.57	3.78	0.98	1.69	21.37	0.98	1.68	90.00
Plackett	1.13	2.26	5.81	1.12	2.17	5.52	1.11	2.40	29.02	1.06	2.56	395.89
Frank	1.18	1.89	3.73	1.13	1.88	4.05	1.11	2.00	17.20	1.06	2.18	215.11
No 16	1.33	4.51	19.15	1.32	4.27	18.48	1.29	4.79	236.85	1.49	4.97	1144.33

In the first instance, it is worth mentioning that the probabilities of failure found in this document are in the same order of magnitude to those obtained by Tang et al. (2013), but they are not equal. There are minor discrepancies due to the modifications applied; specifically, the use of Kendall's tau coefficient instead of Pearson's rho coefficient to compute the copulas dependence parameter contributes significantly to these differences. It is emphasized again that the use of Pearson's rho coefficient must be avoided as a measure to characterize the dependence in copula theory. Rank measures of dependence are more suitable for this purpose. Furthermore, when the dataset is available, the methods exposed in section 4.4.4 are also appropriated to fit copulas to data.

Nonetheless, it is worth noting that for the same dataset, and for the same copula, methods exposed



**Figure 4-18:** Probabilities of failure associated with different copulas for an infinite slope analysis. (a)  $H$  varying from 6.27m to 3.84m, (b)  $\alpha$  varying from  $32^\circ$  to  $27.7^\circ$ , (c)  $\lambda_{c,\phi}$  varying from 0.67 to 1.64, and (d)  $\tau_{c,\phi}$  varying from  $-0.26$  to  $-0.78$

in section 4.4.4 may also lead to a different copula dependence parameters, although quite similar. These minor discrepancies in the dependence parameter of a given copula will result in different probabilities of failure. Further studies are necessary to explore the repercussions of the copula fitting methods on the probability of failure of geotechnical structures.

Regarding the results presented in figure 4-18 and table 4-15, it is important to present some

conclusions that help to understand the implications of using copulas when performing reliability analysis in geotechnical engineering. More details and remarks in this regard are presented by [Tang et al. \(2013\)](#) and [Tang et al. \(2015\)](#) for the infinite slope example, or by [Phoon and Ching \(2014\)](#), [Xu et al. \(2016a\)](#), [Li et al. \(2015b\)](#) and [Xu and Zhou \(2018\)](#) over other structures and conditions. Thus:

- As it was expected, as  $H$  or  $\alpha$  decreases, the probability of failure becomes smaller, and consequently, the factor of safety increases. This conclusion is easily extendable to other geotechnical models by saying that as the geometric parameters become more adverse for the stability of the geotechnical system, the probability of failure will be greater, and vice versa.
- As the coefficient of variation C.O.V. decreases (or  $\lambda_{c,\phi}$  increases), the probability of failure obtained through all the copulas becomes smaller. This behavior makes sense since a decrease in the C.O.V. means a decrement of the standard deviation (keeping the mean unchanged), and this leads to less heavy-tailed multivariate distributions.
- As Kendall's tau coefficient approaches to  $-1$ , the probability of failure decreases. This behavior is expected since the closer the dependence approaches perfection, the less dispersion is in the final distribution, which leads to smaller probabilities of failure.
- For the same geotechnical model and dependence value between random variables, different copulas lead to different probabilities of failure. Discrepancies among their results are more pronounced when the reliability of the model is high, i.e. for small probabilities of failure. In contrast, it is seen that probabilities of failure obtained from different copulas approach to a same value when the reliability of the model is low, i.e. for high probabilities of failure.
- Differences among probabilities of failure are more sensitive to changes in statistical parameters than in geometric parameters. That is to say, the highest discrepancies among probabilities of failure are found when the C.O.V., or the dependence parameter  $\tau$  of the random variables change.
- Note from figure 4-18 that the Gaussian copula leads to the smallest probabilities of failure for all the evaluated conditions. Different studies agree that the lowest probabilities of failure, or the highest reliability indexes ([Zhang et al., 2014](#)), are always obtained from the Gaussian copula for each particular geotechnical model analyzed. This fact indicates that the use of this copula, or the commonly employed multivariate normal distribution, overestimates the reliability, and hence, it may be too unconservative.
- Among all the studied conditions and parameters, changes in Kendall's tau are the ones that lead to the highest discrepancies in the probability of failure obtained from different copulas. This fact shows the importance of correctly fitting the copulas since the probability of failure is quite sensitive to this parameter.

- Note that all the studied copulas can approach both a perfect negative dependence and independence. Thus, when  $\tau$  approaches to 0, all the obtained probabilities of failure should converge to the same value, corresponding to the one obtained from the independence copula  $\Pi(u, v)$ . On the other hand, when  $\tau$  approaches to  $-1$ , all probabilities of failure should approach the same value, given by the countermonotonicity copula  $W(u, v)$ . Although this behavior was not explicitly evidenced in this document, it was pointed out by other investigations, such as [Tang et al. \(2015\)](#).
- Although both Student- $t$  and Gaussian copulas are elliptical copulas, their probabilities of failure differ significantly, especially in high reliability levels. These discrepancies are mainly due to the tail dependence of the Student- $t$  copula. In this way, tail dependence plays an important role when performing reliability analyses, since the computed probabilities of failure may vary in several orders of magnitude by taking it into account or not.
- In this particular infinite slope example, the MCS performs well in comparison to the direct integration method applied by [Tang et al. \(2013\)](#). However, the true benefits of the MCS over the direct integration method appear when the performance function is highly non-linear or when there are more than two random variables involved in the reliability analysis. In this way, the MCS is a more general methodology than the direct integration method, and its use is recommended for the computation of probabilities of failure.

Finally, it is worth mentioning that some studies such as [Wu \(2013b\)](#), [Wu \(2015\)](#), or [Zhang et al. \(2014\)](#), prefer to express the reliability in terms of reliability indexes instead of probabilities of failure. Particularly, [Zhang et al. \(2014\)](#) noted that it is preferable to employ the reliability index when performing a weighting of the results, in order to avoid disproportional effects given by the large differences between probabilities of failure. Unlike probabilities of failure of different copulas, where the discrepancies between them may be of several orders of magnitude, differences in reliability indexes are not as pronounced. Commonly, all the reliability indexes of a geotechnical model obtained through different copulas are in the same order of magnitude, which makes their discrepancies minor.

### **4.7.3 Efforts to unify results of reliability analysis of geotechnical structures when dependence is considered.**

In the last section, it was demonstrated through an infinite slope example that different copulas lead to different values of probabilities of failure. Particularly, for a case of very low probabilities of failure, the results obtained from different copulas may differ by several orders of magnitude. Based on this, the question that arises is which is the true or real reliability of the geotechnical system? or which of all these probabilities of failure better represent the reliability?.

Firstly, one could think about selecting the probability of failure given by the copula that best fits data, which in turn is chosen through one of the methods of section 4.4.5. However, although

this is a valid option, it was demonstrated in section 4.6 that defining the best-fit copula could be misleading, especially when no sufficient data is available. In other words, there is not a complete certainty of choosing the best-fit copula when the amount of data is small, which is quite common in geotechnical engineering practice.

As noted by Zhang et al. (2014), there would not be model selection uncertainty if the probability of one model being the real one were 1.0, while the probabilities of the others were zero. This fact seldom occurs, but what is certain is that there are models more likely to be the correct ones in contrast to the others. Nonetheless, it can not be said under these circumstances that these high-likely models have the final word regarding the probability of failure since the inherent model uncertainty would be denied.

A good estimative of the uncertainty when selecting models is through the ratio  $P(M_1|D)/P(M_2|D)$ , where  $P(M|D)$  is the probability of model  $M$  being the true one, given the data  $D$ . If this ratio is equal or larger than 20, the hypothesis of  $M_1$  being overwhelmingly better than  $M_2$  is correct. On the contrary, if this ratio is between 1.0 and 3.0 this hypothesis becomes only a vague evidence in favor of  $M_1$  (see, e.g., Kass and Raftery, 1995).

The previous estimative was employed by Zhang et al. (2014). In this work several bivariate copula models were fitted to a shear strength dataset, and each model probability was computed based on a bayesian approach. As expected, some copula models (with their respective marginal distributions) give a better fit to data than other copulas. However, no model overwhelmingly fits better with data in comparison to the others. In particular, the ratio between the probabilities of the most likely models is just 1.1, which indicates that model uncertainties can not be neglected.

In order to characterize the dispersion between the probabilities of failure obtained through different copulas, the global dispersion factor  $r$  (see Dutfoy and Lebrun, 2009), was employed by Tang et al. (2015) in an example of an infinite homogeneous slope. The formula of the global dispersion factor is given by:

$$r = \frac{P_{f_{max}}}{P_{f_{min}}}$$

where  $P_{f_{max}} = \max \{P_f(C)\}$ , and  $P_{f_{min}} = \min \{P_f(C)\}$ , where  $C \in \mathcal{C}$  is a copula function,  $\mathcal{C}$  is the subset of all the studied copulas, and  $p_f(C)$  is the probability of failure obtained through the copula  $C$ .

A small  $r$  indicates that the probabilities of failure vary in a narrow range, while a large  $r$  means a substantial dispersion between the probabilities of failure. As stated by Tang et al. (2015), if  $1.0 \leq r \leq 1.5$ , the probability of failure can be evaluated quantitatively. On the other hand, if  $1.5 < r \leq 10.0$ , the probability of failure can be evaluated qualitatively. And finally, if  $r > 10.0$ , then the estimated probability of failure may exceed the real probability of failure by at least one order of magnitude, which may be too inaccurate and dangerous when defining the reliability of a geotechnical structure.

Other good estimate of the dispersion in the results obtained from different copulas is the local dispersion factor  $r_0$ . The local dispersion factor refers to the relation between the  $P_{f_{max}}$  (or  $P_{f_{min}}$ ) and the probability obtained from certain copula  $C$  in  $C$ . Its formula is expressed as follows:

$$r_0 = \max \left\{ \frac{P_{f_{max}}}{P_f(C)}, \frac{P_f(C)}{P_{f_{min}}} \right\}$$

Note that the global dispersion factor quantifies the maximum dispersion in the probabilities of failure obtained from the copulas of  $C$ , whereas the local dispersion factor quantifies the maximum dispersion concerning a reference copula in  $C$ .

It is interesting to compute and plot the global dispersion factors from the homogeneous infinite slope example of section 4.7.2. This procedure was performed and figure 4-19 presents a curve for each condition studied in the last section that represents the global dispersion factor.

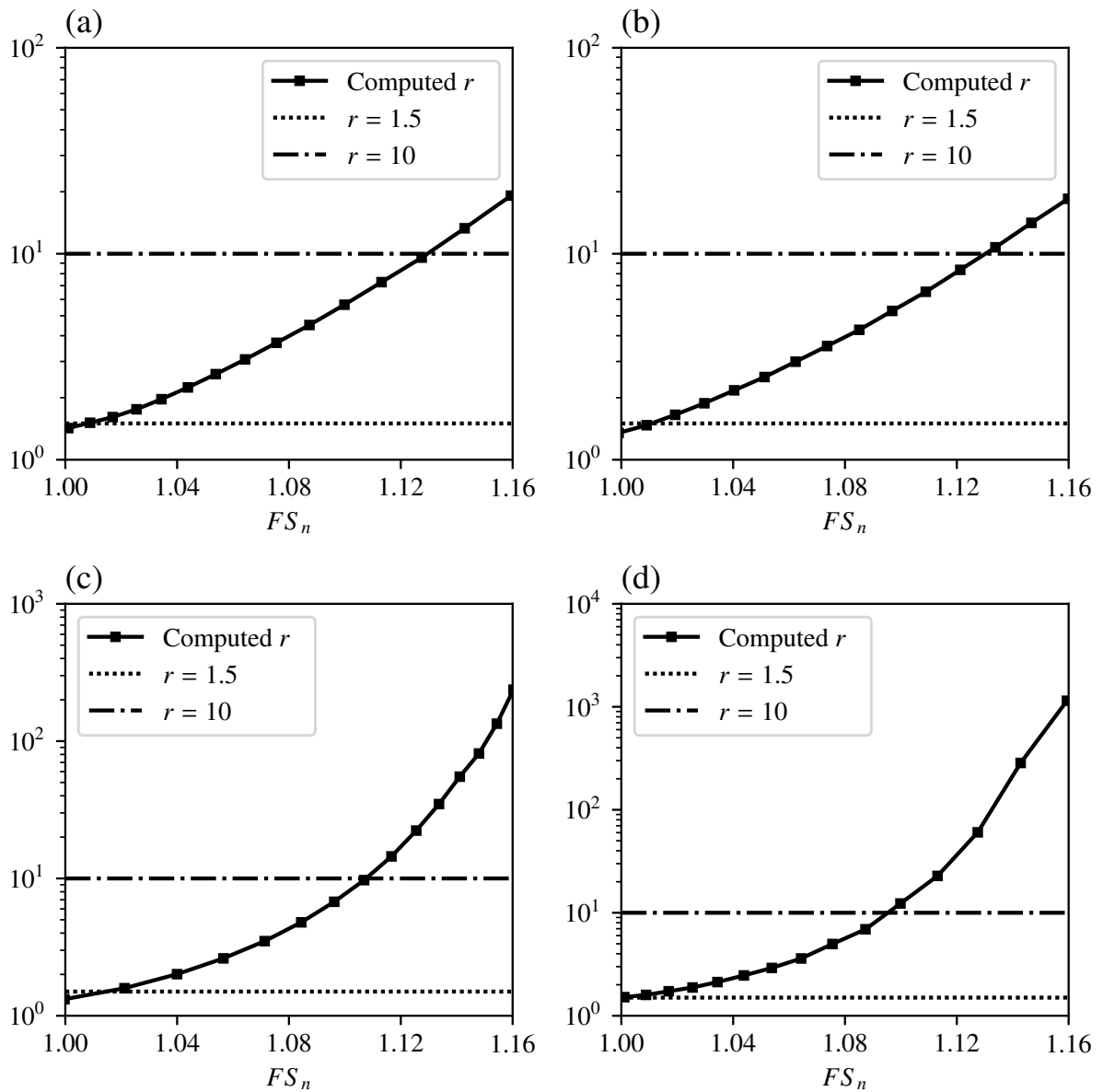
As exposed previously, differences among the probabilities of failure obtained through each copula (i.e.,  $P_f(C)$ ), are more sensitive to changes in  $\tau$  or  $\lambda$  than to changes in geometric parameters. This behavior is evident in figure 4-19, since global dispersion factors are greater for these probabilistic parameters, and particularly, they are greater for changes in  $\tau$ .

Now, it is interesting to compare, with help of the figure 4-19, the obtained global dispersion factors with the intervals defined by Tang et al. (2015). Following this guideline, it is found that for all the studied conditions the probability of failure can be evaluated quantitatively just for the smallest factor of safety, close to 1.0. For larger factors of safety ( $\approx 1.02$  or higher) the probability of failure must be evaluated qualitatively. And, for the highest factors of safety ( $\approx 1.12$  or higher), the estimated probability of failure may exceed the real probability by at least one order of magnitude. These comparisons show great dispersions between results obtained through different copulas.

The above justifies the need for a more consistent estimative of the probability of failure, which has to take into account the inherent uncertainty. For this purpose, several methodologies have been employed in the literature on copula theory in geotechnical engineering. Specifically, we found interesting three methodologies presented by Tang et al. (2015) to overcome this issue, which in turn are based on the postulates of Dutfoy and Lebrun (2009).

Taking into account that the three methodologies exposed by Tang et al. (2015) serve to understand other methodologies and approaches that complement them, these will be briefly presented below. Subsequently, comparisons are made with the approaches and modifications presented in other studies, such as Zhang et al. (2014) or Xu et al. (2016a). Thus, the three methodologies of Tang et al. (2015) are summarized as follows:

1. The first approach consists of, conservatively, taking the copula in  $C$  which leads to the highest estimative of the probability of failure. It is common to see this practice in geotechnical engineering, especially when the amount of data is scarce, and hence there is high uncertainty. However, its major disadvantage is that it could be too conservative, which increases



**Figure 4-19:** Global dispersion factors of the studied infinite slope in section 4.7.2. (a)  $H$  varying from 6.27 to 3.84, (b)  $\alpha$  varying from  $32^\circ$  to  $27.7^\circ$ , (c)  $\lambda_{c,\phi}$  varying from 0.67 to 1.64, and (d)  $\tau_{c,\phi}$  varying from  $-0.26$  to  $-0.78$

the robustness of the designs/solutions, and hence their price significantly.

2. The second approach is based on the local dispersion factor defined by the equation 4.7.3. Briefly, the copula belonging to  $C$  that leads to the smallest local dispersion factor must be selected, and hence, its probability of failure must be adopted. This approach minimizes local dispersions in the final probability of failure and provides more reasonable estimates.

3. The third approach, and perhaps the most appealing, consists of a weighted average of all the copulas belonging to  $C$ . With this approach, a new copula is constructed in the form of:

$$C(u, v; \theta) = \sum_{i=1}^n p_i C_i(u, v, \theta_i) \quad (4-42)$$

where  $p_i$  is the probability of each copula to be the correct one, so  $\sum_{i=1}^n p_i = 1.0$ , and  $C_i$  is the copula  $i$  in  $C$ , with its respective fitting parameters represented by the vector  $\theta_i$ .

This approach has the advantage of being straightforward but, at the same time, robust. It allows taking into account every copula of  $C$  and to make estimates of, e.g., the probability of failure of the model. In the final estimates, each candidate copula has a certain incidence that may be greater or less, in concordance with the respective probability of being the correct copula.

Regarding the probability of each copula being the correct one ( $p_i$  in the equation (4-42)), there have been exposed several methodologies in the literature to address the computation of this value. For example, the procedure presented by [Tang et al. \(2015\)](#) for this purpose is based on a bootstrap approach, quite similar to what was done in section 4.6. In this sense, [Tang et al. \(2015\)](#) employed the shear strength dataset, reported by [Tang et al. \(2013\)](#), to conduct a bootstrap procedure. A total of ten thousand bootstrap samples ( $N_s = 10000$ ) were recreated by [Tang et al. \(2015\)](#), and for each one, the AIC value of the different copulas was computed. Finally, the number of times that each copula is defined as the best fit copula is divided by the number of bootstrap samples ( $N_z$ ), which leads to the probability of each copula being the correct one ( $p_i$ ).

Another way to compute  $p_i$  values from equation (4-42) was presented by [Zhang et al. \(2014\)](#), which in turn was applied over a shear strength dataset to demonstrate the procedure. In this work, the Bayes' theorem, based on a set of observed data ( $D$ ), is employed to calculate the posterior probability of copula model  $M_i$  being the correct one, as follows:

$$P(M_i|D) = \frac{P(D|M_i) P(M_i)}{\sum_{j=1}^r P(D|M_j) P(M_j)} \quad (4-43)$$

where  $P(M_i|D)$  is posterior probability of  $M_i$  being the true model,  $P(D|M_i)$  is the probability of observe  $D$  given that the model is  $M_i$ , and  $P(M_i)$  is the prior probability that the model  $M_i$  is the true one. If no prior probability is specified, then it could be assumed that  $P(M_i) = \frac{1}{r}$ , where  $r$  is the number of candidate models in  $C$ .

Note that  $P(M_i|D)$  in equation (4-43) is the same  $p_i$  of equation (4-42). Additionally, after some approximations, and based on [Kass and Raftery \(1995\)](#), equation 4-43 can be rewritten as follows (see [Zhang et al., 2014](#)):

$$P(M_i|D) \approx \frac{\exp[-\Delta_i(BIC)/2]}{\sum_{j=1}^r \exp[-\Delta_j(BIC)/2]}$$

in which

$$\Delta_i(BIC) = BIC_i - \min_{j=1,2,\dots,r} \{BIC_j\}$$

where  $BIC$  is the Bayesian Information Criterion exposed in section 4.4.5.2. This approximation makes it easier to compute the probability of each model ( $p_i$ ), which makes this method as straightforward and easily-applicable, just like the bootstrap approach employed by Tang et al. (2015).

Another approach to compute  $p_i$  that is worth mentioning is the entropy weights (Zou et al., 2006). This approach was applied in geotechnics by Xu et al. (2016a) to weigh the results of different copulas, applying equation (4-42), when calculating the probability of failure of a slope with a rotational failure mechanism. Note that all the aforementioned weighting methods vary only in the manner of computing  $p_i$ . Thus, the principle by itself is equal, since the calculated  $p_i$  are replaced in the equation (4-42) to weight the results obtained through all the copulas.

In this way, equation (4-42) has been applied by several studies to weigh results obtained from different copulas, specifically when these are employed to conduct geotechnical reliability analysis. The main difference of some studies when applying equation (4-42) is the way in which these compute the probability of each model to be the correct one  $p_i$ .

In conclusion, from our point of view the second and third methods, presented by Tang et al. (2015), are the most adequate to take into account the model uncertainty in the estimates of the final probability of failure. Between these two methods, the one based on the weighted average may be more appealing than the one based on the local dispersion factor. The first method presented by Tang et al. (2015), although it is a common practice in geotechnical engineering, it is also too conservative and must be avoided whenever it is possible.

However, even though the aforementioned approaches serve to take into account the model uncertainties, and certainly these compute more robust estimates of the probability of failure, they are not perfect methodologies. In fact, more investigation is needed in this regard, since model uncertainty, especially when the amount of data is small, is still a challenging problem.

# 5 Random sets and evidence theory in geotechnical engineering

So far, two key developments to geotechnical reliability analyses have been postulated. Firstly, the subset simulation algorithm allows obtaining an accurate estimate of the probability of failure of a geotechnical model, being computationally efficient but keeping the robustness of the MCS. Secondly, copula theory allows to evaluate and include dependence among geotechnical parameters. This dependence is commonly obviated or assumed without grounds, but as demonstrated before, it has many implications in the final reliability estimates. The issue now is how to include epistemic uncertainty in reliability analyses, specifically in geotechnical engineering where this type of uncertainty has such a significant impact.

In our practice, it is usual to have incomplete and insufficient information, which makes it impossible to define with complete certainty the limit state function  $G$  and the joint PDF  $f_X$  of the models. Under these circumstances, the purely probabilistic approach that assesses reliability through equation (2-5) lose applicability (Blockley, 1980, 1999). Let us keep in mind that imprecise information cannot give precise conclusions, and hence, the final probability of failure is not known with total certainty (Bernardini and Tonon, 2010). For instance, Oberguggenberger and Fellin (2002, 2004) demonstrated that the probability of failure of a shallow foundation may vary in several orders of magnitude when using different likely probability distributions associated with soil parameters, even if these distributions are defined based on a relatively large number of samples. Furthermore, sometimes it is even impossible to gather field or laboratory data, and the only available information is the judgement and opinion of local geotechnical professionals, which is commonly expressed in terms of intervals or linguistic terms<sup>1</sup>.

There exists several techniques to represent uncertainty in practical applications (see, e.g., Joslyn and Booker, 2004, for an overview of these approaches). Traditionally, probability theory has been the standard approach to deal with aleatory uncertainty. But on the other hand, several approaches have lately emerged to address the issue of epistemic uncertainty. However, when both aleatory and epistemic uncertainties are present in an engineering system, some of these methodologies treat to convert one type of uncertainty into another. This procedure carries out several assumptions which are mostly not justified, and hence it may lead to mistakes.

---

<sup>1</sup>For instance, Rackwitz (2000) presents interval ranges for the means and standard deviations of some soil properties from several common types of cohesive and non-cohesive soils.

Fortunately, among all these approaches, *random sets theory* stands out for being a theory that manages to integrate several methodologies in order to create an unifying framework that allows working with aleatory and epistemic uncertainties without any assumption. Random sets theory emerges from the independent works of [Kendall \(1974\)](#) and [Matheron \(1974\)](#), who studied the behavior of random variables valued in closed bounded sets of  $\mathbb{R}^d$ . This approach can be seen as an extension of the point-valued probability theory to a set-valued probability theory. Furthermore, this approach is also capable of integrating several structures of uncertainty representation such as intervals, possibility distributions, probability boxes and, of course, probability functions (all these structures will be further explained later).

Therefore, this chapter introduces random sets theory and its potential role when conducting geotechnical reliability analysis considering both aleatory and epistemic uncertainties. This chapter starts elucidating *Dempster-Shafer evidence theory*, which is closely related to random sets theory. Later, random sets theory is introduced, and also its relationship with several structures for representing uncertainty. Finally, the propagation of both epistemic and aleatory uncertainty through an engineering model is explained. The chapter ends with a practical example on the use of random sets theory in geotechnical engineering. After finalizing this chapter, the importance of epistemic uncertainty will be evident, especially in the particular case of geotechnical engineering, where information is scarce.

## 5.1 Dempster-Shafer evidence theory

Dempster-Shafer evidence theory (DST), also known as evidence theory, is a mathematical theory for representing both aleatory and epistemic uncertainties. The early work conducted by [Dempster \(1967\)](#) was later expanded in the seminal work of [Shafer \(1976\)](#) to what is known today as Dempster-Shafer evidence theory. In essence, this theory can be seen as an extension of the conventional probability formulation, where probabilities are assigned to sets as opposed to mutually exclusive singletons ([Sentz et al., 2002](#)). In other words, while traditional probability theory associates evidence with one possible event, DST is capable to associate evidence with multiple possible events, e.g., sets of events. In this sense, as a generalization of probability theory, Dempster-Shafer theory collapses to the traditional probabilistic formulation when the evidence is sufficient enough to assign probabilities to single events. Thus, DST allows a direct representation of both epistemic and aleatory uncertainties since the methodology is designed to cope with varying levels of precision and information.

Following the definition of Dempster-Shafer structures given by [Dubois and Prade \(1991\)](#), let us consider an universal non-empty set  $X$ , and its power set  $\mathcal{P}(X)$ . Furthermore, let

$$\mathcal{F} := \{A_1, A_2, \dots, A_n\}$$

be a family of several different non-empty sets of  $X$ , such that  $A_i \in \mathcal{P}(X) \setminus \emptyset$  and  $A_i \neq A_j$  for all

$i, j = 1, 2, \dots, n$  and  $i \neq j$ . Also let

$$m : \mathcal{F} \rightarrow [0, 1]$$

be a mapping from  $\mathcal{F}$  to the unit interval.

Every set  $A_i \in \mathcal{F}$  is called a *focal element*, and all of them have an associated *basic mass assignment*  $m(A_i) \geq 0$ . Furthermore, the collection of focal elements  $A_i$  is known as the *focal set*  $\mathcal{F}$ . Thus, the tuple  $(\mathcal{F}, m)$  is known as a *Dempster-Shafer body of evidence*, or simply a *Dempster-Shafer structure* on  $X$ . Furthermore, some researches have coined the term *evidence space* for the tripla  $(X, \mathcal{F}, m)$  (e.g., Helton et al., 2006). Note that every  $A_i$  contains the possible values of the variable  $x \in X$  and  $m(A_i)$  is the probability of  $A_i$  being the true range of  $x$ . In this order of ideas, the probability mass assignment has the following properties:

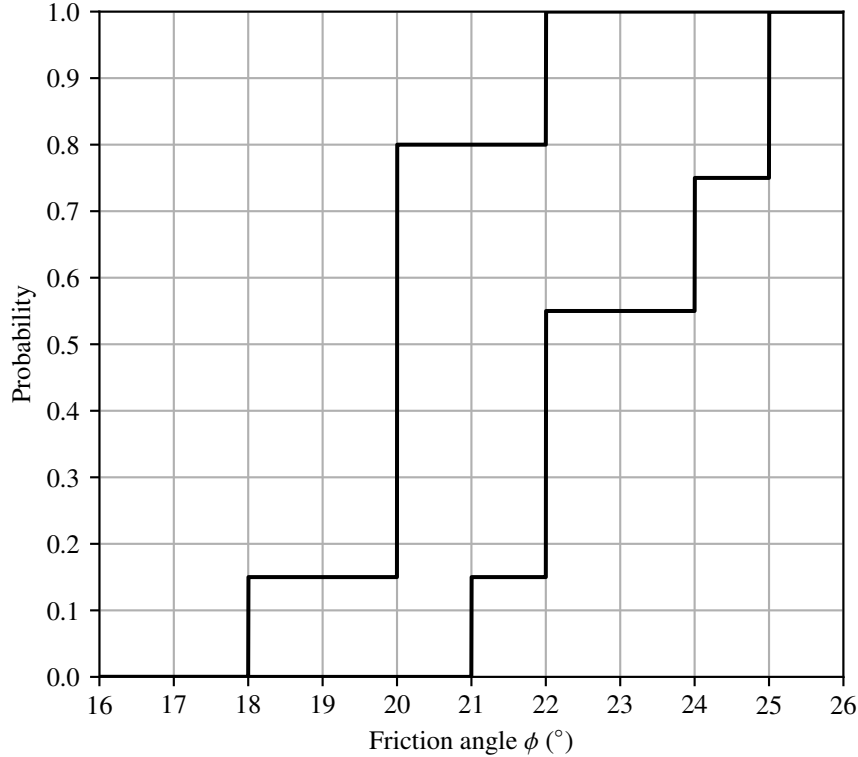
$$m : \mathcal{F} \rightarrow [0, 1]$$

$$m(\emptyset) = 0$$

$$\sum_{A \in \mathcal{F}} m(A) = 1$$

It is worth mentioning that the probability mass assignment of  $A_i$  expresses the extend to which the available information supports the claim that a particular element of  $X$  belongs to  $A_i$ , but it says nothing about the element being within the subsets of  $A_i$  (Klir, 1995). Any further information about the particular element being within a subset of  $A_i$  must be represented by another focal element and its respective mass assignment (Bernardini and Tonon, 2010). To illustrate the above concepts, let us consider the following toy example.

**Example 5.1.1.** (adapted from Alvarez, 2007) Let us suppose we are interested in defining the inner friction angle of some soil, and for this purpose we gather information from different sources (e.g., books, experts, previous analysis, etc.). Employing these sources of information, we define four focal elements  $A_i$ , which constitute our focal set  $\mathcal{F}$ . For this hypothetical case,  $\mathcal{F} := \{A_1 := [20^\circ, 22^\circ], A_2 := [22^\circ, 24^\circ], A_3 := [18^\circ, 21^\circ], A_4 := [20^\circ, 25^\circ]\}$ . Furthermore, we set the basic mass assignment for every element of our focal set, in concordance to the evidence of the investigation and the importance of the sources of information in the assessments. In this example,  $m(A_1) = 0.4$ ,  $m(A_2) = 0.20$ ,  $m(A_3) = 0.15$ , and  $m(A_4) = 0.25$ . Then, the tuple  $(\mathcal{F}, m)$  is constituted and it defines our Dempster-Shafer structure. Note that some focal elements share information (e.g.,  $A_1$  and  $A_4$ ) and some others contains conflicting information (e.g.,  $A_2$  and  $A_3$ ). The issue of conflicting information will play a key role when combining different sources of information (Sentz et al., 2002). Figure 5-1 illustrates the Dempster-Shafer structure obtained in this example.



**Figure 5-1:** Illustration of the Dempster-Shafer structure from example 5.1.1

Note that from our hypothetical information gathered, the soil friction angle from this example is enclosed by two bounds, as represented in figure 5-1. In other words, it is expected that any realization of a given value of the friction angle will be within these two bounds, and never outside of them. The concept of these two measures will be explained below.

### 5.1.1 Belief and Plausibility measures

Dempster-Shafer theory states that, in the light of imprecise information, the probability that an element  $x$  belongs to a set  $F$ , such that  $F \in \mathcal{F}$ , cannot be expressed by a single value, but rather it is bounded by two thresholds, as follows:

$$\text{Bel}_{(\mathcal{F},m)}(F) \leq P(F) \leq \text{Pl}_{(\mathcal{F},m)}(F)$$

where  $\text{Bel}_{(\mathcal{F},m)}(F)$  is called the *belief function*, and  $\text{Pl}_{(\mathcal{F},m)}(F)$  is called the *plausibility function*. These two measures are defined respectively by:

$$\text{Bel}_{(\mathcal{F},m)}(F) := \sum_{i=1}^n \mathbb{I}[A_i \subseteq F] m(A_i) \quad (5-1)$$

and

$$\text{Pl}_{(\mathcal{F},m)}(F) := \sum_{i=1}^n \mathbb{I}[A_i \cap F \neq \emptyset] m(A_i) \quad (5-2)$$

for all  $A_i \in \mathcal{F}$ .

In this way, the belief measure of  $F$  is the sum of all the basic mass assignments of those focal elements contained in  $F$  whose realization implies that  $x \in F$ . On the other hand, the plausibility measure of  $F$  is the sum of all the basic mass assignments of those focal elements that share some element in common with  $F$ , and whose realization may imply that  $x \in F$ . Note that when  $\text{Bel}_{(\mathcal{F},m)}(F) = \text{Pl}_{(\mathcal{F},m)}(F)$ , then DS theory collapses to the conventional probability theory (Alvarez, 2007).

The *degree of ignorance* about the statement of  $x \in F$  is expressed as the difference between the plausibility and belief measures, that is to say, by the quantity  $\text{Pl}_{(\mathcal{F},m)}(F) - \text{Bel}_{(\mathcal{F},m)}(F)$ . This gap represents the epistemic uncertainty, and it will decrease as more and better information is available.

When  $X$  is a finite set, then the basic mass assignment of  $F$  can be restored from the belief or plausibility measured using the Möbius transform (Bernardini and Tonon, 2010):

$$m(F) = \sum_{i=1}^{|\mathcal{F}|} \mathbb{I}[A_i \subseteq F] (-1)^{|A_i \setminus F|} \text{Bel}_{(\mathcal{F},m)}(A_i)$$

Furthermore, the belief and plausibility measures on  $F$  can be related as follows (Wang and Klir, 1992):

$$\text{Bel}_{(\mathcal{F},m)}(F) = 1 - \text{Pl}_{(\mathcal{F},m)}(F^c)$$

$$\text{Pl}_{(\mathcal{F},m)}(F) = 1 - \text{Bel}_{(\mathcal{F},m)}(F^c)$$

and the following properties are met:

- *Super-additivity*:  $\text{Bel}(A \cup B) \geq \text{Bel}(A) + \text{Bel}(B) - \text{Bel}(A \cap B)$
- *Sub-additivity*:  $\text{Pl}(A \cup B) \geq \text{Pl}(A) + \text{Pl}(B) - \text{Pl}(A \cap B)$

Therefore, unlike probability theory where the additivity property  $P(A) + P(A^c) = 1$  holds, in Dempster-Shafer evidence theory  $\text{Bel}(A) + \text{Bel}(A^c) \leq 1$  and  $\text{Pl}(A) + \text{Pl}(A^c) \geq 1$ .

**Example 5.1.2.** Based on example 5.1.1, what would be the probability of the inner friction angle lying in the interval  $F = [22^\circ, 23^\circ]$ ? By using the DST, it could be easily defined that  $\text{Bel}_{(\mathcal{F},m)}(F) = 0.25$  and  $\text{Pl}_{(\mathcal{F},m)}(F) = 0.4 + 0.2 + 0.25 = 0.85$ . In this way, the probability of the friction angle being within  $[22^\circ, 23^\circ]$  is bounded by a probability between  $[0.25, 0.85]$ .

## 5.2 Succinct introduction to random sets theory

This section presents a brief introduction to random sets theory, mostly based on the works of [Alvarez \(2007\)](#), [Bernardini and Tonon \(2010\)](#) and [Bertoluzza et al. \(2002\)](#). So, readers interested in delving into these concepts can consult the aforementioned works and the references contained therein.

### 5.2.1 From random variables to random sets

Let us recall that a random variable is a function that maps the sample space into the measurable space, i.e.,  $X : \Omega \rightarrow \mathbb{R}$ . This transformation allows to mathematically describe a random experiment. Furthermore, note that this mapping associates any event  $\omega \in \Omega$  to a singleton (point)  $x \in \mathbb{R}$  (see figure 2-1).

However, in the presence of epistemic uncertainty a single point in  $\mathbb{R}$  may not be enough to represent an event of  $\Omega$ . Fortunately, random sets theory manages to be an alternative by generating this mapping including both epistemic and aleatory uncertainties.

For RS definition, let us consider an universal non-empty set  $\mathcal{X}$  and its respective power set  $\mathcal{P}(\mathcal{X})$ . Additionally, let  $(\Omega, \sigma_\Omega)$  be a probability space and  $(\mathcal{F}, \sigma_{\mathcal{F}})$  be a measurable space. Then, a random set  $\Gamma$  is a function whose domain is defined on  $\Omega$  and its range on  $\mathcal{F}$ , i.e.,  $\Gamma : \Omega \rightarrow \mathcal{F}$ . As in the conventional probability theory, this mapping serves to generate a probability measure on  $(\mathcal{F}, \sigma_{\mathcal{F}})$ , such that  $P_\Gamma := P_\Omega \circ \Gamma^{-1}$ . In this way, an event  $F$  has the probability:

$$P_\Gamma(F) = P_\Omega(\{\omega : \Gamma(\omega) \in F\}) \quad (5-3)$$

In other words, a random set is a set-valued random variable. To understand these concepts, figures 5-2 and 2-1 can be compared. Furthermore, note that when all elements of  $\mathcal{F}$  are singletons (no epistemic uncertainty), then the random set becomes a random variable, i.e.,  $\Gamma(\omega) = X(\omega)$ , and in this case  $\mathcal{F}$  is said to be *specific*.

Random sets theory states that, in presence of epistemic uncertainty it is not possible to know the exact value of  $P_X(F)$  but its upper and lower bounds. These upper and lower probabilities are defined as ([Dempster, 1967](#)):

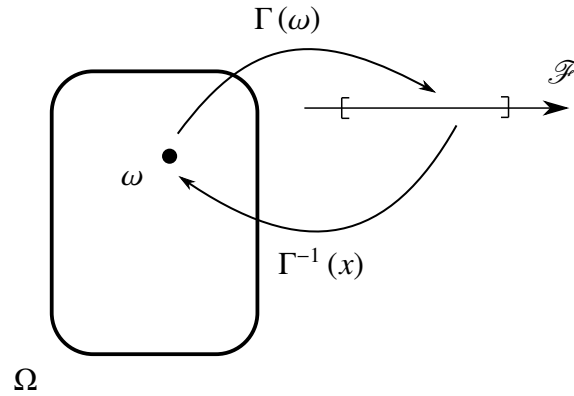
$$LP_{(\mathcal{F}, P_\Gamma)}(F) := P_\Omega(\{\omega : \Gamma(\omega) \subseteq F, \Gamma(\omega) \neq \emptyset\}) \quad (5-4)$$

and

$$UP_{(\mathcal{F}, P_\Gamma)}(F) := P_\Omega(\{\omega : \Gamma(\omega) \cap F \neq \emptyset\}) \quad (5-5)$$

In this sense,  $P_X(F)$  is bounded by the following inequality:

$$LP_{(\mathcal{F}, P_\Gamma)}(F) \leq P_X(F) \leq UP_{(\mathcal{F}, P_\Gamma)}(F)$$



**Figure 5-2:** random set mapping from  $\Omega$  into  $\mathcal{F}$  (adapted from [Tonon et al., 2000a](#))

and when the equality holds, then  $\mathcal{F}$  is specific and the random set collapses to a random variable. Furthermore, note that equations (5-4) and (5-5) are a generalization of the belief and plausibility measures (equations (5-1) and (5-2), respectively) defined by [Dempster \(1967\)](#) for Dempster-Shafer theory.

In this way, random sets theory is closely related to Dempster-Shafer evidence theory. Specifically, when the cardinality of the evidence is finite, Dempster-Shafer structures are mathematically isomorphic to (finite) random sets, although with somewhat different semantics ([Alvarez, 2007](#)). In other words, given a body of evidence  $(\mathcal{F}, m)$  where  $\mathcal{F}_n = \{A_1, A_2, \dots, A_n\}$ , and a random set  $\Gamma : \Omega \rightarrow \mathcal{F}$ , then the following relationship holds:  $\mathcal{F} \equiv \mathcal{F}_n$ , i.e.  $A_j = \Gamma(\omega_j)$ , and  $m(A_j) \equiv P_\Gamma(\Gamma(\omega_j))$ , for  $j = 1, 2, \dots, n$ .

## 5.2.2 Some representations of the state of knowledge and their relationship with random sets theory

Dempster-Shafer structures were previously introduced as a tool for representing epistemic and aleatory uncertainties of a quantity of interest. Furthermore, its relationship with random sets theory was also exposed. Nonetheless, the state of knowledge may not be necessarily represented by a DS but by other structures of a similar nature but somewhat different. Among all these representations we are interested in those with a practical application and that can be generalized by random sets theory. In particular, families of intervals, possibility distributions, probability boxes, and CDFs will be introduced since they may have a wide application in geotechnics. Additionally, the relationships of these structures with RS theory will be elucidated.

### 5.2.2.1 Possibility distributions

A possibility distribution, or a normalized fuzzy set, is a mapping  $A$  from a set  $X$  to the unit interval, i.e.,  $A : X \rightarrow [0, 1]$ , where  $\sup_{x \in X} A(x) = 1$ . Here,  $A(x)$  for  $x \in X$ , represents the degree in which the concept represented by  $A$  is compatible with  $x$ . Furthermore, the  $\alpha$ -cut of a membership function is depicted by the so-called *crisp set*  $A_\alpha = \{x \in X : A(x) \geq \alpha\}$  where  $\alpha \in [0, 1]$ .

Now, let us consider a measurable space  $(\mathcal{F}, \sigma_{\mathcal{F}})$ , in which  $\mathcal{F} \subseteq \mathcal{P}(X)$ . If there exists a family of subsets  $\mathcal{C}_B := \{C : B \subseteq C \in \mathcal{F}\}$  for every  $B \in \mathcal{F}$ , where  $\mathcal{C}_B \in \sigma_{\mathcal{F}}$ , then the function

$$c_\Gamma(B) := P_\Gamma(\mathcal{C}_B) = P_\Omega\{\omega : \Gamma(\omega) \in \mathcal{C}_B\} = P_\Omega\{\omega : B \subseteq \Gamma(\omega)\} \quad (5-6)$$

constitutes a measure on  $X$  for every  $B \subseteq X$  called the *subset coverage function*. It is worth noting that when  $B = \{x\}$ , then  $\mathcal{C}_{\{x\}} = \{C : x \in C \in \mathcal{F}\}$ , and hence equation (5-6) becomes  $c_\Gamma(x) = P_\Omega\{\omega : x \in \Gamma(\omega)\}$  for every  $x \in X$ , such that it constitutes the so-called *one point coverage function* of the random set  $\Gamma$ . Furthermore, note that based on equation (5-5),  $c_\Gamma(x) = UP(\{x\})$ .

Let  $A$  be a possibility distribution on  $X$ , and let  $\tilde{\alpha} : \Omega \rightarrow (0, 1]$  be a uniformly distributed random variable for some probability space  $(\Omega, \sigma_\Omega, P)$ , that is to say,  $P\{\alpha : \tilde{\alpha}(\alpha) \leq z\} = z$  for  $z \in (0, 1]$ . In this way,  $\tilde{\alpha}$  induces a random set  $\Gamma_A(\alpha) = \{x \in X : A(x) \geq \tilde{\alpha}(\alpha)\}$ , which is equivalent to the randomized  $\alpha$ -cut set  $A_{\tilde{\alpha}(\alpha)}$ . Therefore, this is the approach to represent a possibility distribution using a random set (see [Goodman and Nguyen, 2002](#)). The associated random set  $\Gamma$  will have the following one point coverage function:

$$\begin{aligned} c_{\Gamma_A}(x) &= P\{\alpha : x \in \Gamma_A(\alpha)\} \\ &= P\{\alpha : A(x) \geq \tilde{\alpha}(\alpha)\} = A(x) \end{aligned} \quad (5-7)$$

In other words, the random set  $\Gamma_A$  in  $X$  has a point coverage function that coincides with the possibility distribution  $A$ . The *possibility measure*  $\text{Pos}_A : X \rightarrow [0, 1]$  is defined by the one point coverage function (equation (5-7)), as follows:

$$\begin{aligned} \text{Pos}_A(K) &:= \sup_{x \in K} \{c_{\Gamma_A}(x)\} \\ &= \sup_{x \in K} \{A(x)\} \end{aligned}$$

for  $K \subseteq X$ . Furthermore, let us recall that the *necessity measure*  $\text{Nec}_A$  is defined by  $\text{Nec}_A(K) := 1 - \text{Pos}_A(K^c)$ .

Note that the generated random set is nested, in the sense that  $\Gamma_A = \{A_\alpha : 0 < \alpha \leq 1\}$  is totally ordered by set inclusion and that the respective possibility distribution is unimodal.

Based on this, [Dubois and Prade \(1991\)](#) stated that a possibility distribution  $A : X \rightarrow [0, 1]$  can be inner approximated by a finite nested random set  $(\mathcal{F}_n, m)$  of  $n$  focal sets. Thus, let us consider a set

$\{\alpha_1 = 1, \alpha_2, \alpha_3, \dots, \alpha_{n+1} = 0 : \alpha_i > \alpha_j, i > j\}$  of  $\alpha$ -cut levels, then the nested random set  $(\mathcal{F}_n, m)$  defined by:

$$\mathcal{F} = \{A_{\alpha_i} \mid i = 1, \dots, n\}$$

and

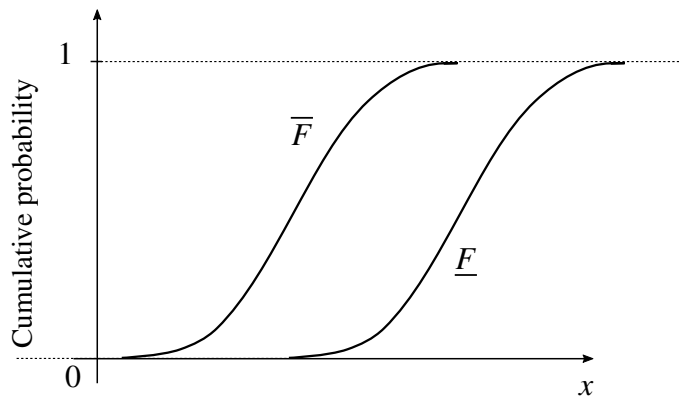
$$m(A_{\alpha_i}) = \begin{cases} \alpha_i - \alpha_{i-1} & i = 1, \dots, n-1 \\ \alpha_n & i = n \end{cases}$$

where  $A_{\alpha_i} := \{x : A(x) \geq \alpha_i\}$  is the  $\alpha$ -level cut, is equivalent to the possibility distribution  $A(x)$ . Commonly, the  $\alpha$ -levels are regularly distributed, and hence,  $m(A_i) = 1/n$ . In an analogous way, the outer approximation can be defined.

### 5.2.2.2 Probability boxes

Let us consider two non-decreasing functions,  $\bar{F}$  and  $\underline{F}$ , that map the real line into the unit interval, i.e.,  $\bar{F}, \underline{F} : \mathbb{R} \rightarrow [0, 1]$ , and that fulfill  $\underline{F}(x) \leq \bar{F}(x)$  for all  $x \in \mathbb{R}$ . Now, let  $[\bar{F}, \underline{F}]$  be the set of all non-decreasing functions  $F$ , where  $F : \mathbb{R} \rightarrow [0, 1]$ , such that  $\underline{F}(x) \leq F(x) \leq \bar{F}(x)$ . Thus, when the pair of functions  $\bar{F}$  and  $\underline{F}$  circumscribe an imprecisely known probability distribution, then  $[\bar{F}, \underline{F}]$  is called a *probability box* or simply a *p-box* (Ferson, 2002; Ferson et al., 2003).

Informally speaking, a p-box groups all the possible distributions  $F$ , that in light of lack of knowledge, can represent a random variable  $X$ . In this sense, although the exact distribution of  $X$  can not be defined, it is known that it falls within  $[\bar{F}, \underline{F}]$ . Hence,  $\underline{F}(x)$  and  $\bar{F}(x)$  are the lower and upper bounds of  $F(x)$ , respectively. Note that the more and better information obtained, the narrower the space between both bounds.  $\bar{F}$  and  $\underline{F}$  will converge to the random variable real distribution as the quantity and quality of information improves. Figure 5-3 presents a scheme of a p-box.



**Figure 5-3:** Scheme of a probability box (adapted from Ferson et al., 2003).

P-boxes may be derived from a variety of states of incomplete knowledge about a quantity, and hence, there exist several ways to construct these structures from analytical judgement and data.

According to [Crespo et al. \(2013\)](#), p-boxes can be classified in two major and disjoint categories: the distributional and the distribution-free p-boxes, as explained below.

In a *distributional p-box*, also known as a parametric p-box, the shape (distribution) of the random variable is known with certainty, but not its parameters. Thus, although the parental distribution is known, its parameters can only be specified by intervals. In this way, the p-box is obtained by enveloping extreme distributions based on the possible parameters. For instance, if the liquid limit of certain soil is known to follow a lognormal PDF with a mean ranging between [5%, 10%] and a standard deviation ranging between [0.1%, 0.3%], then its probability box can be generated based on this information; observe that it represents both aleatory and epistemic uncertainty of the parameter.

On the other hand, in a *distribution-free p-box*, also known as a non-parametric p-box, the parental family of the distribution that specifies the p-box is unknown. In this case, it is required that the CDF of the bounds,  $\bar{F}$  and  $\underline{F}$ , are clearly and explicitly stated, without assumption of the parental distribution. In this sense, distribution-free p-boxes do not have information about the internal structure within the bounds  $\bar{F}$  and  $\underline{F}$ , and hence, they do not state which distribution is the most likely. When additional information is available, then the p-box can be tightened.

In the literature, there is a wide variety of methods for constructing p-boxes from empirical and theoretical information. These methods range from direct assumptions, modeling, robust Bayes approaches, constraint propagation, and observation of measurements. Further details about all the approaches for constructing p-boxes are out of the scope of this document, but readers are referred to [Ferson et al. \(2003\)](#), and the references therein. The choice of the most appropriate method will depend on the type of information available on the parameter of interest. For instance, if only a few samples from a large population are available, then experimental measurements methods may be a right choice, and particularly, the Kolmogorov-Smirnov (K-S) confidence limits may be employed ([Kolmogoroff, 1941](#); [Smirnov, 1939](#)). An example of this construction is presented below:

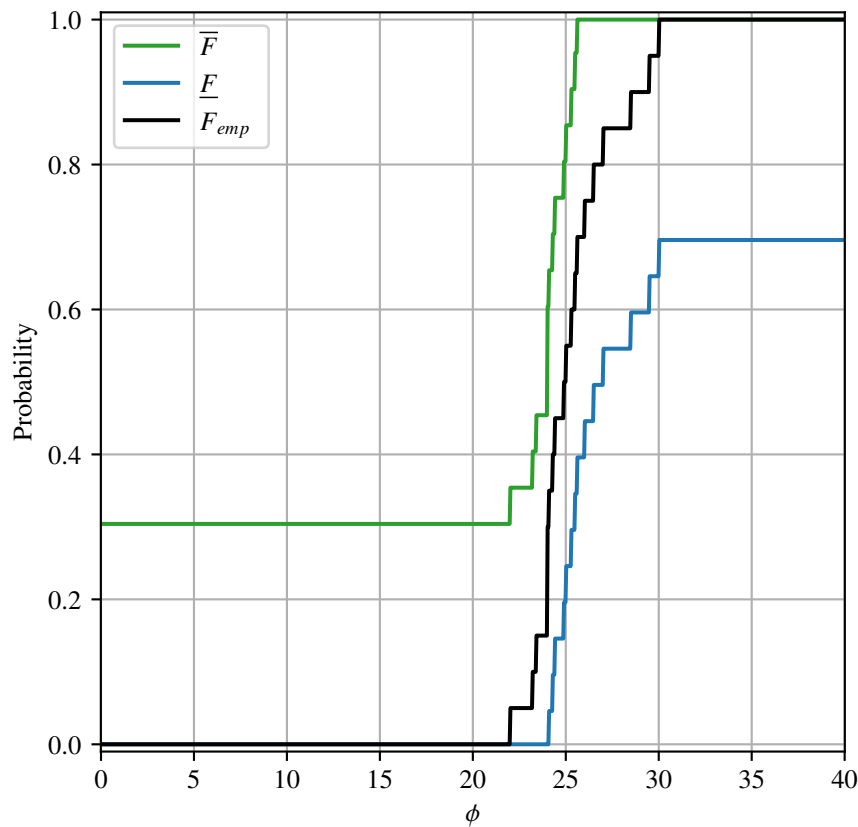
**Example 5.2.1.** (adapted from [Alvarez, 2007](#)) Let us consider the following dataset consisting of twenty inner friction angles obtained from shear tests over the same soil and under the same conditions ([Oberuggenberger and Fellin, 2005, 2008](#)).

$$\phi (^{\circ}) = \begin{cases} 22.0, 23.2, 23.4, 24.0, 24.0, 24.0, 24.1, 24.3, 24.4, 24.9, \\ 25.0, 25.3, 25.5, 25.6, 26.0, 26.5, 27.0, 28.5, 29.5, 30.0 \end{cases}$$

In the light of conventional probability approach, the common practice would be to choose a CDF, that fulfills some goodness-of-fit hypothesis test (e.g., the Kolmogorov-Smirnov test, the Anderson-Darling test, etc.), for representing the dataset. Nonetheless, a single CDF would deny the inherent epistemic uncertainty of the model, which is quite significant when the available information is scarce.

On the other hand, from this dataset one can employ the Kolmogorov-Smirnov confidence limits

methodology to construct a p-box that takes into account both aleatory and epistemic uncertainty. Figure 5-4 shows the results of this procedure for a confidence level of 90%, provided that the samples are independent and identically distributed. Furthermore, figure 5-4 also shows the empirical CDF for the dataset of this example.



**Figure 5-4:** Empirical CDF ( $F_{emp}$ ) and the probability box  $[\bar{F}, F]$  for  $\phi$ , obtained by using the K-S test for a confidence level of 90% and the dataset from the example 5.2.1

Observe that by using the K-S confidence limits methodology one obtains a distribution-free p-box about an empirical CDF. Furthermore, it is worth mentioning that the bounds from a K-S test are not certain bounds, but statistical ones. In other words, the associated statistical statement is that 90% (or any other confidence level) of the time the true distribution will lie inside these bounds.

Now, the issue is how to associate probability boxes with random sets theory. In fact, there is a close relationship between both structures. Every RS generates an unique p-box, and every p-box generates an equivalent class of random intervals. More specifically, every p-box can be discretized to obtain a finite random set (see, e.g., Ferson et al., 2003). In this regard, Alvarez (2006) showed

that a distribution-free p-box can be related to a random set as follows:

$$\Gamma(\omega) = \left[ \overline{F}_x^{-1}(\omega), \underline{F}_x^{-1}(\omega) \right] \quad (5-8)$$

for  $\omega \in \Omega$ .

However, equation (5-8) is not directly applicable to distributional p-boxes. These kinds of p-boxes must be first converted into a distribution-free p-box in order to specify a RS through equation (5-8). This is mainly due to the inherent inability of RS theory to take into account the parental CDF that defines the distributional p-box. This transformation may lead to a loss of information. Nonetheless, [Alvarez et al. \(2017\)](#) explains how to deal with distributional p-boxes without losing information.

### 5.2.2.3 Families of intervals

An interval  $I = [a, b]$  is a range of values between a lower limit  $a$  and an upper limit  $b$ . This structure do not involve any knowledge about the CDF or shape of the values between the two limits, and hence, all CDFs are possible. In this sense, an interval should not be confused with an uniform CDF in  $[a, b]$ , since the former is an expression of the lack of knowledge and the latter is a representation of randomness.

Intervals are widely employed in geotechnical engineering practice. Many books and references present intervals for different parameters of common soil/rock materials. Furthermore, it is quite common that when asking a local geotechnician about some parameter, she/he expresses it in terms of an interval or in linguistic terms that can be later transformed into intervals.

The relationship between intervals and random sets is straightforward. A single interval can be seen as a RS with an unique focal element  $A$  with  $m(A) = 1$ . When there is a family of equally probable  $n$  intervals, each one of these is considered to be a focal element  $A_i$  with a corresponding  $m(A_i) = 1/n$ . Furthermore, if the intervals of a family are not equiprobable, then a Dempster-Shafer body of evidence can be constituted and directly related to a random set.

### 5.2.2.4 Probability functions

A probability function can be approximated by a histogram with  $n$  discrete intervals. Each one of these intervals constitutes a focal element of a random set, and the height of each interval in the histogram constitutes the basic mass assignment. It is worth noting that when the probability function has an unbounded domain (e.g., normal distribution), it is required to impose upper and lower bounds in order to stablish the RS. These bound, for instance, can be taken as the 0.005 and 0.995 percentiles of the probability function ([Alvarez, 2007](#)).

### 5.2.3 Combination of evidence

In some situations, it may be possible to have multiple bodies of evidence, obtained from different sources, for the same uncertain quantity of interest. For instance, two geotechnicians can have different opinions about some soil property, say  $(\mathcal{F}_A, m_A)$  and  $(\mathcal{F}_B, m_B)$ , respectively. Under these circumstances, it may be necessary to mix all these sources into a new body of evidence  $(\mathcal{F}_{joint}, m_{joint})$  that combines all the opinions.

However, this procedure is not a trivial task and there are multiple strategies reported in the literature to face this labor. A comprehensive exposure of several of these methods, and their applicability, is given by [Sentz et al. \(2002\)](#). Furthermore, [Ferson et al. \(2003\)](#) also explained some of these techniques and gave some remarks and guidelines on choosing the most suitable approach as a function of particular conditions. Nonetheless, the combination of evidence is still an open and unsolved problem, especially when there are large discrepancies between the different bodies of evidence ([Sentz et al., 2002](#)).

Among all the strategies that exist to combine evidence, the *Dempster's rule of combination* ([Dempster, 1967](#); [Shafer, 1976](#)) is worth mentioning, since many other approaches are derived from it. Dempster rule states that bodies of evidence can be combined as follows:

$$\mathcal{F}_{joint} := \mathcal{F}_A \cup \mathcal{F}_B \quad (5-9)$$

and

$$m_{joint}(A_i) := \frac{\sum_{A \cap B = A_i} m_A(A) m_B(B)}{1 - K} \quad (5-10)$$

where

$$K = \sum_{A \cap B \neq \emptyset} m_A(A) m_B(B)$$

for all  $A_i \neq \emptyset$ ,  $A \in \mathcal{F}_A$ ,  $B \in \mathcal{F}_B$  and  $m_{joint}(\emptyset) = 0$ . Note from equation (5-10) that  $K$  accounts for the contradictory of conflicting information among the sources. The normalization factor  $(1 - K)$  on the denominator of equation (5-10) has the effect of completely ignoring conflicting evidence and attributing any mass assignment associated with conflict to the null set ([Yager, 1987](#)). Therefore, one of the disadvantages of Dempster rule is that it loses its applicability when the degree contradiction between sources of information significantly increases.

### 5.2.4 Random relations for random sets

Random relations for random sets is a concept that appears when dealing with a function of several variables. These functions require to associate the different variables of the model in order to conduct the analyses. But, how different random sets can be associated?

To answer this question, let  $X := \times_{i=1}^d X_i$  be a multivariate space in  $\mathbb{R}^d$ . Thus, a random relation for a finite RS  $(\mathcal{F}, m)$  is defined as the cartesian product on  $X$  given by the combination of the marginal randoms set  $(\mathcal{F}^i, m^i)$ , where  $\mathcal{F}^i = \{A_{j_i}^i : j_i = 1, \dots, n_i\}$  for  $i = 1, \dots, d$ , such that:

$$(\mathcal{F}, m) := (A_{j_1, \dots, j_d} := \times_{i=1}^d A_{j_i}^i, m_{j_1, \dots, j_d} := f(m_{j_1}^1, \dots, m_{j_d}^d)). \quad (5-11)$$

From equation (5-11) it is worth mentioning that the function  $f$  is the multivariate probability function that associates the basic mass assignments of the different variables. In this sense, this function takes into account the dependence relation between the marginal random set. For instance, if there is certainty that the variables are independent, then random set independence can be employed (Oberkampf et al., 2004), and hence the basic mass assignment for the multivariate RS can be stated as  $m(A_{j_1, \dots, j_d}) := \prod_{i=1}^d m^i(A_{j_i}^i)$  for all  $A_{j_1, \dots, j_d} \in \mathcal{F}$ .

However, as stated in chapter 4, dependence plays a key role when propagating uncertainty, and so in the final computations and reliability estimates. Therefore, it is not good practice to make assumptions when there is not enough evidence that supports some claims about the dependence relationship. In this regard, when nothing is known about the dependence between variables, unknown interaction may be a conservative but suitable choice (Fetz and Oberguggenberger, 2004). More about dependence in random set will be further faced in the subsequent section.

### 5.2.5 Dependence considerations

Chapter 4 of this document was entirely dedicated to the subject of dependence in geotechnical engineering, and for this purpose, copula theory was introduced and its role when modeling relationships between different variables explained. However, although this theory opens a wide range of possibilities for modeling dependence, it is a fact that it is difficult to define the best copula function for a set of parameters when information is scarce. In this order of ideas, there is also uncertainty when defining the best copula for modeling any dependence in geotechnical engineering, and the degree of uncertainty will depend mostly on the availability of information for defining the structure of dependence.

Uncertainty in the definition of dependence for a set of parameters can also be addressed from random set theory. Fetz and Oberguggenberger (2004) and Couso et al. (1999) defined several kinds of dependence from the perspective of RS theory, which result from the combination of the marginal random sets. Furthermore, Oberkampf et al. (2004) expose the role of dependence when dealing with p-boxes and Dempster-Shafer structures, and also multiple conditions of dependence were introduced. A common conclusion from these works is that dependences should not be assumed so lightly, and the usual assumption of independence can be dangerous when conducting reliability analysis. This conclusion agrees with what was evidenced in chapter 4 of this thesis.

It is out of the scope of this document to enter in details of the kinds of dependences defined by Fetz and Oberguggenberger (2004) and Couso et al. (1999), but it is worth mentioning some of

these structures, particularly *random set independence* and *unknown interaction*. Therefore, in the remaining of this document we will consider an arbitrary dependence for the combination of those marginal focal elements. In this sense, one should work with the better fit copula defined by the methods exposed in chapter 4. More about the employability of copulas in p-boxes and DSS can be found in [Oberkampf et al. \(2004\)](#).

The concept of random set independence states that for a given random set  $(\mathcal{F}, m) := \times_{i=1}^d (\mathcal{F}^i, m^i)$  the basic mass assignments are defined as  $m(A_{j_1, j_2, \dots, j_d}) := \prod_{i=1}^d m^i(A_{j_i}^i)$ , while dependence selections are allowed for the probability measures inside  $A_{j_1, j_2, \dots, j_d}$ . It is worth mentioning that the criteria for employing random set independence must be quite well founded since otherwise its assumption may be harmful.

On the other hand, the concept of unknown interaction appears when nothing is known about dependence for assigning the basic mass assignments to the set  $A_{j_1, j_2, \dots, j_d}$ . In this criteria the probability measures on  $A_{j_1, j_2, \dots, j_d}$  are allowed to vary freely between the Frechet-Hoeffding bounds, i.e. between perfect dependence and perfect opposite dependence. [Fetz and Oberguggenberger \(2004\)](#) developed a linear programming approach for estimating the bounds of  $m$  when unknown interaction in random set is employed.

It is worth mentioning that when more information about dependence is available, the range in which it varies could be narrowed. For example, this would be the case of physical conditions that make certain dependence values impossible. Another case would be when it is known with complete certainty that the dependence is positive or negative. Finally, another example could be when it is known for sure that the dependence range is between two limits that are narrower than the Frechet-Hoeffding bounds. A complete discussion on this issue is provided by [Oberkampf et al. \(2004\)](#) and the references cited therein.

## 5.2.6 Extension principle for random sets

After establishing the variables of interest and defining the structures that best represent them, e.g., probability boxes, intervals, probability functions, etc., the issue is then how to propagate them through the models. In other words, given the function  $G : X \rightarrow Y$  and the random set  $(\mathcal{F}, m)$ , one may be interested in the image that  $(\mathcal{F}, m)$  projects through  $G$ , i.e.  $(\mathcal{R}, \rho)$ . This concept is useful, for example, for knowing how uncertainties are propagated through the limit state function  $G$  and how they affect the final reliability estimate.

This mapping can be carried out by the application of the extension principle, such as follows ([Dubois and Prade, 1991](#)):

$$\begin{aligned} \mathcal{R} &:= \{R_j := G(A_i) : A_i \in \mathcal{F}\} \\ \rho(R_j) &:= \sum_{i=1}^n \mathbb{I}[R_j = G(A_i)] m(A_i) \end{aligned} \tag{5-12}$$

Conventionally,  $X \subseteq \mathbb{R}^d$ , and hence every  $A_i \in \mathcal{F}$  corresponds to a  $d$ -dimensional box with  $2^d$  vertices, obtained applying the cartesian product explained in section 5.2.4. Furthermore, note that equation (5-12) requires the definition of the image that a focal element  $A_i$  projects through  $G$ . Thus, it is required to solve two global optimization problems, such that:

$$G(A_i) = [l_i, r_i]$$

where

$$l_i = \min_{u \in A_i} G(u)$$

$$r_i = \max_{u \in A_i} G(u)$$

The process of propagating the focal elements through the function  $G$ , i.e. conducting equation (5-12), when  $Y \subseteq \mathbb{R}$ , can be performed by one of the following methodologies: the optimization approach, the sampling approach, the vertex approach, and the function approximation approach (Alvarez, 2007).

### 5.2.6.1 Method of optimization

The optimization method consists in performing a numerical optimization using the values  $x$  within the focal element  $A_i$ , finding thus the values that minimize and maximize the performance function, i.e.  $[\min_{x \in A_i} G(x), \max_{x \in A_i} G(x)]$ . This procedure is applicable when the focal elements  $A_i$  are compact and the performance function  $G$  is continuous. Furthermore, this approach is quite useful when the performance function  $G$  is a nonlinear function of the system parameters. However, the main drawback of this methodology is its complexity and the excessive computational effort required for complex and large scale systems.

### 5.2.6.2 Method of sampling

The sampling method consists of, as its name states, drawing samples within the domain that constitutes the  $d$ -dimensional box formed by the marginal random set. For this purpose, an arbitrary probability density function can be employed for drawing samples from each focal element (e.g., uniform PDF is a good option given its simplicity). Subsequently, the performance function  $G$  is evaluated for each drawn point, in order to define the belief and plausibility measures that projects the random set through the model.

If the number of samples is large enough, then this method achieves an acceptable approximation to  $\min_{x \in A_i} G(x)$  and  $\max_{x \in A_i} G(x)$ ; otherwise, the approximation may be vague and coarse. In this way, as the number of samples increases, the better the approximation to the belief and plausibility measures. This procedure is quite demanding computationally speaking, and it may even be unfeasible if great accuracy in the evaluation of  $R_j$  is required.

This method is not computationally efficient in high-dimensions because an acceptable estimate of  $R_j$  would require a large number of samples. In this way, the sampling method suffers the *curse of dimensionality* since the number of points required to populate the space increases dramatically as the dimension of the problem increases. Furthermore, when the configuration of the problem is not advantageous (i.e. complex nonlinear performance functions) then the method may also lead to wrong approximations of  $R_j$ .

### 5.2.6.3 Method of vertices

The vertex method (Dong and Shah, 1987) consists of evaluating the performance function only for the vertices of the  $d$ -dimensional box generated by the marginal random set, in order to find the maximum and minimum responses generated by the performance function. Evidently, this methodology is more efficient than the optimization method or the sampling method. However, the application of this approach is only valid when the performance function is monotonic and continuous concerning each variable; otherwise, this method would provide erroneous results. It is worth mentioning that as the discretization of the information structures becomes finer, the approximation of  $R_j$  given by this methodology will be much more accurate (see, e.g., Tonon, 2004). Furthermore, in the case of a non-monotonic performance function, the vertex method can be employed as an approximation to  $R_j$  that gets closer to the true result as the size of the focal elements decreases.

### 5.2.6.4 Method of function approximation

The idea behind the function of approximation (Hurtado and Alvarez, 2000) consists of improving the efficiency of the above methods by fit the performance function  $G$  to a more simple surrogate function  $\hat{G}$ , called the *response surface*. The function  $\hat{G}$  must approximate  $G$  up to some degree of accuracy and also be more tractable computationally than  $G$ . In this way, the efficiency of the above algorithms can be enhanced since there is no need to call the complex function  $G$  but the response surface. Nonetheless, it is worth mentioning that the chosen surrogate function largely affects the accuracy of the algorithm, so it must be selected wisely.

## 5.3 Employability of random sets theory in geotechnical engineering

Random sets theory has been successfully applied in many areas of knowledge, such as big data, artificial intelligence, medicine, natural hazard risk assessments, structural reliability, hydrology, etc. However, although its applicability in geotechnical engineering has great potential, the reality is that its use has been rather scarce. Nonetheless, there are pioneering works that have successfully

used the methodology in soil and rock engineering, in addition to the fact that the interest of geotechnicians in this theory has been increasing in recent times.

Among all the pioneering works that have used random sets theory in geotechnical engineering some of them are worth mentioning. For instance, the early work of [Tonon et al. \(1996\)](#) employed random sets theory to evaluate the response of rock mass within a tunnel excavation. On the other hand, [Tonon et al. \(2000a\)](#) relied on random sets theory to propose a methodology for estimating the rock mass parameters (RMP) of a rock mass and presented some examples of its applicability. Furthermore, [Tonon et al. \(2000b\)](#) employed Dempster-Shafer evidence theory and Monte Carlo simulation to perform some estimations of plane failure and tunnel lining designs.

Other interesting works in this framework integrate random sets theory with the finite element method. This approach is commonly known as random sets finite element method (RS-FEM). For instance, [Peschl \(2004\)](#) developed a methodology in which the inputs of a finite element geotechnical model are a random set, and hence the probability of the model is bounded. Furthermore, later works have also employed RS-FEM in geotechnical engineering, for example, [Schweiger and Peschl \(2004\)](#) in deep excavations, [Schweiger and Peschl \(2005\)](#) in slopes and anchored sheet piles, or [Peschl and Schweiger \(2004\)](#) in geotechnical reliability problems general in practice. This methodology has also been applied to the study of tunnel excavations [Nasekhian and Schweiger \(2011\)](#).

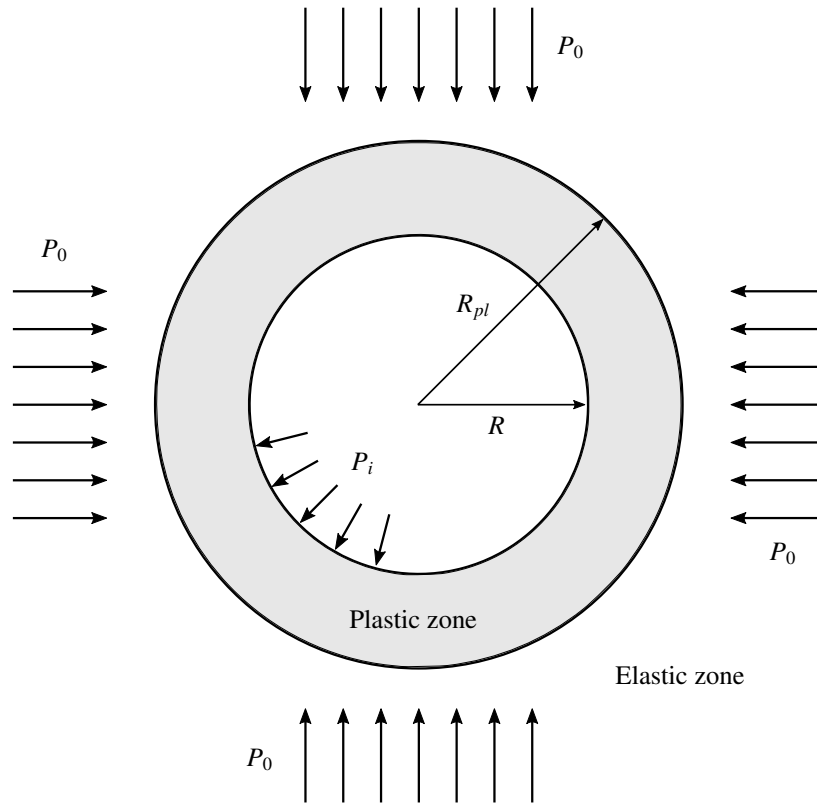
All the aforementioned publications are good examples of the applicability of random sets theory in geotechnical engineering. Therefore, readers are encouraged to consult and study them. Additionally, it is worth mentioning that, although these works have used RS theory in certain particular problems, its applicability has not been widely extended. That is to say, this theory can be employed in many other problems related to soils and rocks, so there is still much to explore regarding its use in geotechnics.

In this document, in order to expose the applicability of random sets theory for propagating uncertainties in a geotechnical model, a practical example is performed. This example consists in the estimation of the convergence of a tunnel, subject to hydrostatic pressure, for which there are both epistemic and aleatory uncertainties in its input parameters that must be propagated through the model.

### 5.3.1 Practical example on the use of random sets in a geotechnical model

Let us consider a circular tunnel excavated in a continuous, homogeneous, and isotropic rock mass, as schematized in figure [5-5](#).

The tunnel has a radius of excavation  $R$ , a radial uniform support pressure of  $p_i$ , and it is subject to a pressure  $p_0$  given by the hydrostatic stress field. With these conditions, the inward displacement



**Figure 5-5:** Circular tunnel in a homogeneous, continuous, and isotropic rock mass, subject to a hydrostatic stress field.

of the tunnel lining  $u_{ip}$  can be computed as follows (see, e.g., [Hoek, 2000](#)):

$$\frac{u_{ip}}{R} = \left[ \frac{1 + \mu}{E} \right] \left[ 2(1 - \mu)(p_0 - p_{cr}) \left( \frac{R_{pl}}{R} \right)^2 - (1 - 2\mu)(p_0 - p_i) \right] \quad (5-13)$$

where  $\mu$  is the Poisson ratio,  $E$  is the Young's modulus, and  $R_{pl}$  is the radius of the plastic zone. The parameters  $p_{cr}$  and  $R_{pl}$  can be computed as:

$$p_{cr} = \frac{2p_0 - \sigma_c}{k + 1}$$

$$\frac{R_{pl}}{R} = \left[ \frac{2(p_0 + s)}{(k + 1)(p_i + s)} \right]^{1/(k-1)}$$

with

$$k = \frac{1 + \sin(\phi)}{1 - \sin(\phi)}$$

$$\sigma_c = \frac{c(k - 1)}{\tan(\phi)}$$

$$s = \frac{\sigma_c}{k - 1}$$

where  $\phi$  and  $c$  are the friction angle and the cohesion of the rock mass, respectively. Furthermore, note that a plastic yielding occurs if  $p_{cr} > p_i$ .

It is known with certainty that  $p_0 = 3.0$  MPa and  $p_i = 0.8$  MPa. Nonetheless, the parameters  $\mu$ ,  $E$ ,  $c$ , and  $\phi$ , have both aleatory and epistemic uncertainties in their definition.

Under these conditions, it is required to estimate the uncertainties in the calculation of the ratio of the inward displacement of the tunnel lining to the tunnel radius, i.e.  $u_{ip}/R$  from equation (5-13). The information available about the uncertain parameters is as follows:

- The four uncertain parameters (i.e.,  $c$ ,  $\phi$ ,  $E$  and  $\mu$ ) are assumed to be independent. That is to say, the knowledge of the value of one parameter implies nothing about the values of the other parameters.
- The Poisson ratio  $\mu$  of the rock mass is given by a triangular probability distribution, with a limit value of  $\mu_{min} = 0.30$ , a mode of  $\mu_{mod} = 0.35$ , and an upper value of  $\mu_{max} = 0.40$ .
- The Young's modulus  $E$  of the rock mass is given by three independent and equally credible sources of information. Each source establishes a closed interval  $E_i$  of possible values that  $E$  can adopt. Thus,  $E_1 = [325 \text{ MPa}, 350 \text{ MPa}]$ ,  $E_2 = [360 \text{ MPa}, 375 \text{ MPa}]$ , and  $E_3 = [395 \text{ MPa}, 395 \text{ MPa}]$
- The cohesion  $c$  of the rock mass counts with three independent and equally credible sources of information for its definition. All three sources agree that  $c$  is given by a triangular distribution; however, each source states different closed intervals in which the fitting parameters of the distribution lie, such as follows:
  - First source:  $c_{min_1} = [0.11 \text{ MPa}, 0.14 \text{ MPa}]$ ,  $c_{mod_1} = [0.15 \text{ MPa}, 0.16 \text{ MPa}]$ ,  $c_{max_1} = [0.18 \text{ MPa}, 0.20 \text{ MPa}]$
  - Second source:  $c_{min_2} = [0.10 \text{ MPa}, 0.13 \text{ MPa}]$ ,  $c_{mod_2} = [0.14 \text{ MPa}, 0.17 \text{ MPa}]$ ,  $c_{max_2} = [0.19 \text{ MPa}, 0.21 \text{ MPa}]$
  - Third source:  $c_{min_3} = [0.09 \text{ MPa}, 0.12 \text{ MPa}]$ ,  $c_{mod_3} = [0.13 \text{ MPa}, 0.17 \text{ MPa}]$ ,  $c_{max_3} = [0.19 \text{ MPa}, 0.22 \text{ MPa}]$

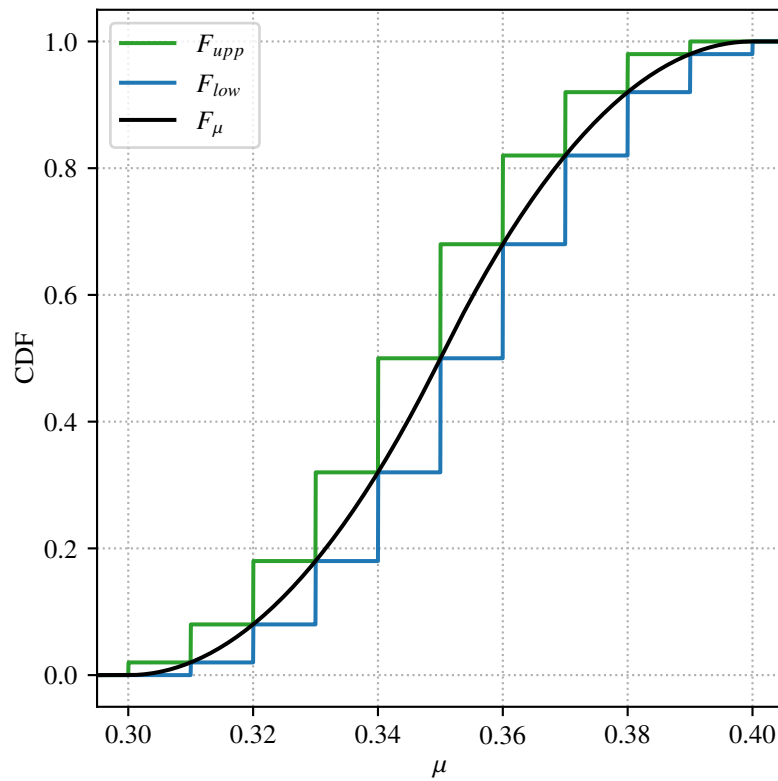
furthermore, for any combination of values, the relationship  $c_{min} \leq c_{mod} \leq c_{max}$  holds.

- The friction angle  $\phi$  of the rock mass is given by a triangular probability distribution whose lower limit value lies in the interval  $[15^\circ, 18^\circ]$ , whose upper limit value lies in the interval  $[25^\circ, 27^\circ]$ , and whose mode lies in the interval  $[20^\circ, 24^\circ]$ . For any combination of values the relationship  $\phi_{min} \leq \phi_{mod} \leq \phi_{max}$  holds.

In order to solve this problem, it is necessary to identify the type of information available for each parameter and relate it to random sets theory. For this purpose, the structure of information of each parameter is discretized (when required) into  $n = 10$  elements. This number of elements is

convenient for the clarity of the graphs. Nonetheless, it is worth mentioning that better results are expected if finer discretizations are employed.

The discretization of each one of the basic variables of the model is explained below, and the resulting structure is also presented graphically. Numerical values of the resulting focal elements, and their respective mass assignments, are not presented nor listed in tables, this for the sake of readability since they are too many elements. Nonetheless, readers can easily perform the discretization and obtain the numerical values of the focal elements, following the procedure outlined.

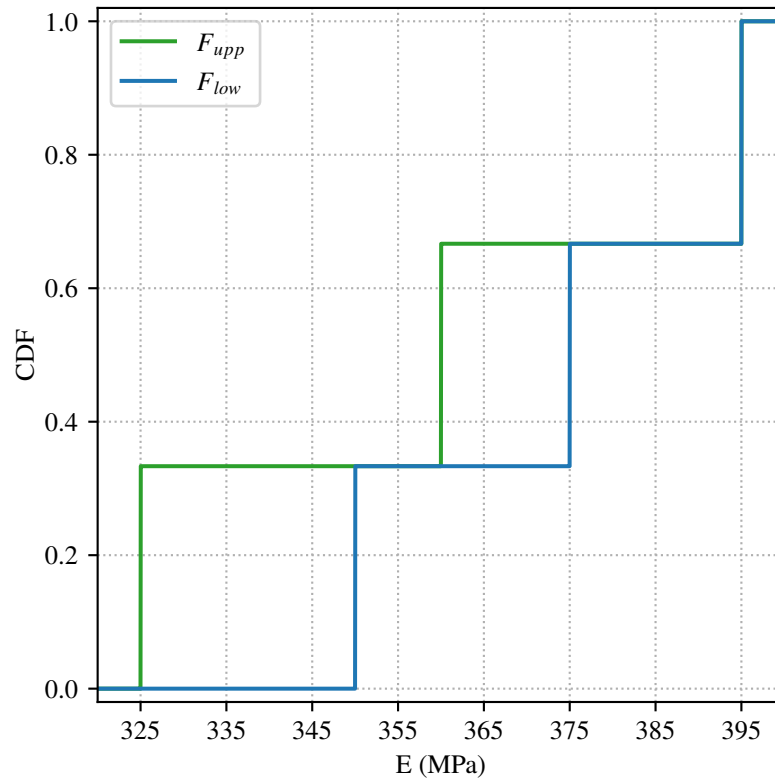


**Figure 5-6:** Example 5.3.1: cumulative distribution function of  $\mu$  ( $F_{\mu}$ ) and its upper CDF  $F_{upper}$  and lower CDF  $F_{lower}$  after the discretization of  $F_{\mu}$  into  $n = 10$  focal element.

The Poisson ratio  $\mu$  has only aleatory uncertainty in its definition, so it is simply modeled as a triangular CDF with the parameters mentioned above. Therefore, the discretization of  $\mu$  into a random set is conducted for computational purposes only. This procedure is straightforward since the CDF of  $\mu$  can be simply divided into  $n = 10$  equally spaced sub-intervals  $A_{\mu,i} = [a_i, b_i]$  (see, e.g., Tonon et al., 2000b). Figure 5-6 presents the final result after the discretization of the CDF.

The Young's modulus  $E$  is affected by both aleatory and epistemic uncertainty. However, since the information is given by intervals, then the relationship with random sets theory is straightforward.

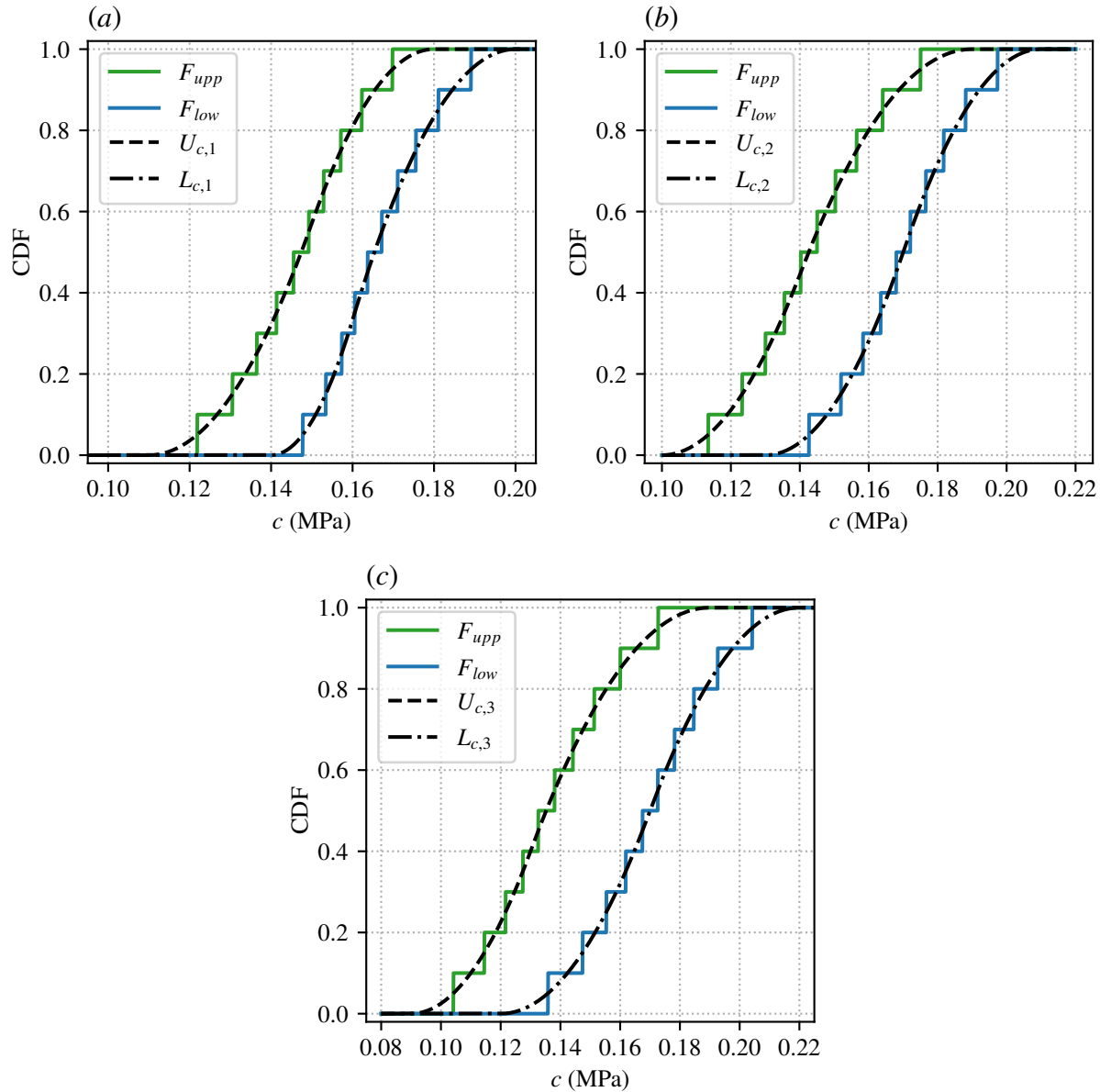
Each interval constitutes a focal element, and since the three sources of information are independent and equally credible, the basic mass assignment for each element is defined just as  $m_E = 1/3$ . Figure 5-7 depicts structure of evidence for the parameter  $E$ .



**Figure 5-7:** Example 5.3.1: Upper ( $F_{upper}$ ) and lower ( $F_{lower}$ ) CDFs of the Young's modulus.

Regarding the cohesion  $c$ , there is information provided by three independent and equally credible bodies of evidence, each of which establishes that the parameter has both aleatory and epistemic uncertainty, and that it is represented by a triangular distribution. In this sense, it is necessary to first define the structure that represents each body of evidence and then combine them all to obtain the final uncertainty structure of  $c$ .

Since each body of evidence assigns compact intervals to the minimum, median and maximum values of the distribution, there is an envelope of triangular CDFs compatible with the given information. Such a set of envelopes are bounded by an upper CDF,  $U_{c,i}$ , and a lower CDF,  $L_{c,i}$ , where the subscript  $c$  refers to the cohesion and the subscript  $i$  refers to the  $i$ -th source of information. The upper bound,  $U_{c,i}$ , can be defined by a triangular distribution in the interval  $[\min_{\min_i}, \max_{\max_i}]$ , with a mode  $\min_{mod_i}$ . Furthermore, the lower bound,  $L_{c,i}$ , can be defined by a triangular distribution in the interval  $[\max_{\min_i}, \max_{\max_i}]$  with a mode  $\max_{mod_i}$ . In this way, an analogous structure to a distributional p-box can be constituted from each source of information.



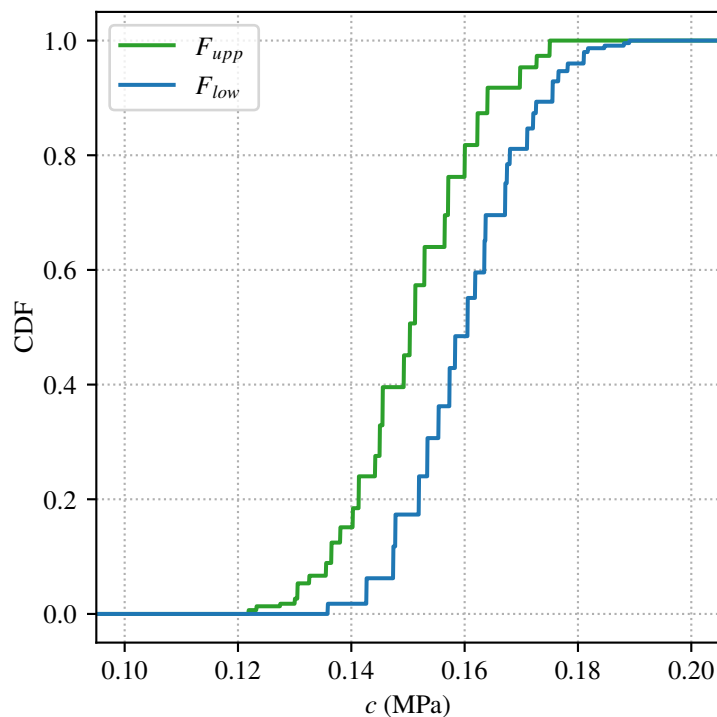
**Figure 5-8:** Example 5.3.1: upper ( $U_{c,i}$ ) and lower ( $L_{c,i}$ ) CDFs of  $c$ , and their discretization into  $F_{upp}$  and  $F_{low}$ , respectively. (a) first source of information, (b) second source of information, and (c) third source of information.

Subsequently, it is required to discretize  $U_{c,i}$  and  $L_{c,i}$  in order to obtain the belief and plausibility functions of an equivalent random set. For this purpose, a total of  $n = 10$  equally probable sub-intervals, i.e.,  $m_k = 1/n$ , are generated employing the Averaging Discretization Method (ADM) (Tonon, 2004, see). This method produces a random set in which its plausibility averages  $U_{c,i}$  and its belief averages  $L_{c,k}$ . Therefore, this approach is quite effective for computing the expectation of a parameter or the expectation of its image through a function. Equations and further details

about the ADM can be found in [Tonon \(2004\)](#). Thus, by following this procedure the discretization depicted in [figure 5-8](#) is obtained.

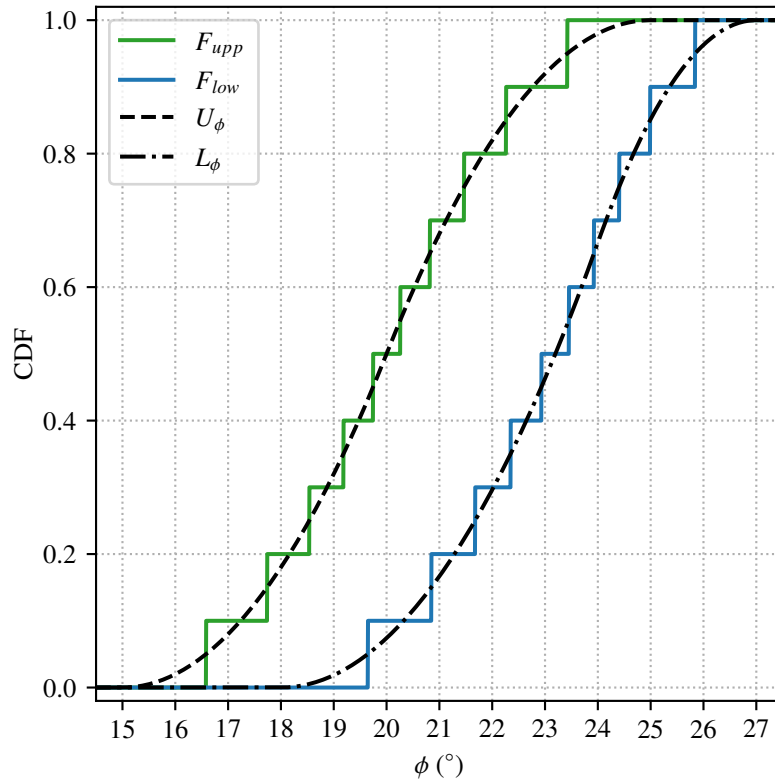
Now, the final step for obtaining the resulting body of evidence for  $c$  is to combine the information given by the three sources. This procedure can be performed by, e.g., employing Dempster's rule. That is to say, the resulting focal elements can be obtained by the equation (5-9), and their respective mass assignments by the equation (5-10). Nonetheless, it is worth highlighting again that, within RS theory, the combination of two or more bodies of evidence is still an open and unsolved problem. Dempster's rule performs well in this example since the conflict between the three sources of information is small; but otherwise, Dempster's rule would lose its applicability. Furthermore, other rules of combination could also be employed in this case. For instance, [Alvarez \(2007\)](#) used the intersection rule for aggregation of p-boxes (see, e.g., [Ferson et al., 2003](#)).

After performing the combination of the three sources of information, a less imprecise body of evidence is obtained. [Figure 5-9](#) presents the  $F_{upp}$  and  $F_{low}$  CDFs of the resulting structure of information. It is worth mentioning that by discretizing each body of evidence in  $n = 10$  focal elements, and after combining them all using Dempster's rule, the final random set contains 146 focal elements.



**Figure 5-9:** Example 5.3.1: upper ( $F_{upp}$ ) and lower ( $F_{low}$ ) CDFs of  $c$  after applying Dempster's rule for combining the initial three sources of information.

Finally, the friction angle  $\phi$  of the rock mass is also affected by aleatory and epistemic uncertainty. It is known with certainty that  $\phi$  follows a triangular CDF, but its adjustment parameters are uncertain and expressed by means of intervals. Therefore, there is an envelope of triangular CDFs compatible with the given information. Note that this is the same case of the parameter  $c$  but considering only one source of information, so it is not necessary to perform the combination of evidence. Thus, following the same procedure described for  $c$ , the upper  $U_\phi$  and lower  $L_\phi$  CDFs are obtained for  $\phi$ , and subsequently, they can be discretized into  $n = 10$  focal elements thus obtaining  $F_{upp}$  and  $F_{low}$ , respectively. Figure 5-10 shows the upper and lower CDFs for  $\phi$ , and also their discretization into  $F_{upp}$  and  $F_{low}$ .



**Figure 5-10:** Example 5.3.1: upper ( $U_\phi$ ) and lower ( $L_\phi$ ) CDFs of the parameter  $\phi$  of the rock mass, and their discretization  $F_{upp}$  and  $F_{low}$ , respectively

At this point, each uncertain parameter of equation (5-13) is constrained by a random set. Therefore, the next step is to define the random relation of them all. For this purpose, the concepts seen in section 5.2.4 are applied, taking into account that all the parameters are assumed to be independent. In our particular case:

$$\mathcal{F}\{A\} = \{A_{\mu,i} \times A_{E,j} \times A_{c,h} \times A_{\phi,l}\}$$

and

$$\begin{aligned} m(A) &= m(A_{\mu,i} \times A_{E,j} \times A_{c,h} \times A_{\phi,l}) \\ &= m_{\mu}(A_{\mu,i}) m_E(A_{E,j}) m_c(A_{c,h}) m_{\phi}(A_{\phi,l}) \end{aligned}$$

for all combinations of  $i, j, h, l$ .

Consequently, the random relation  $(\mathcal{F}, m)$  is formed. Note that each element of  $\mathcal{F}$  is a four-dimensional box with  $2^4$  vertices, since they result from the cartesian product of four intervals.

Given that the random relation now is established, it is then the time to map  $(\mathcal{F}, m)$  to the mechanical system response through function  $u_{ip}/R$  (equation (5-13)). For this purpose, the extension principle has to be applied (see section 5.2.6). In particular, the vertex method is employed here for this purpose, and hence the function of  $u_{ip}/R$  is calculated 16 times for each  $A_i \in \mathcal{F}$  (one for each vertex of the four-dimensional box). Then, for each multidimensional box the local maxima and minima of the function  $u_{ip}/R$  are defined. Hence, the image that  $(\mathcal{F}, m)$  project to  $u_{ip}/R$ , i.e.  $(\mathcal{R}, \rho)$ , can be established.

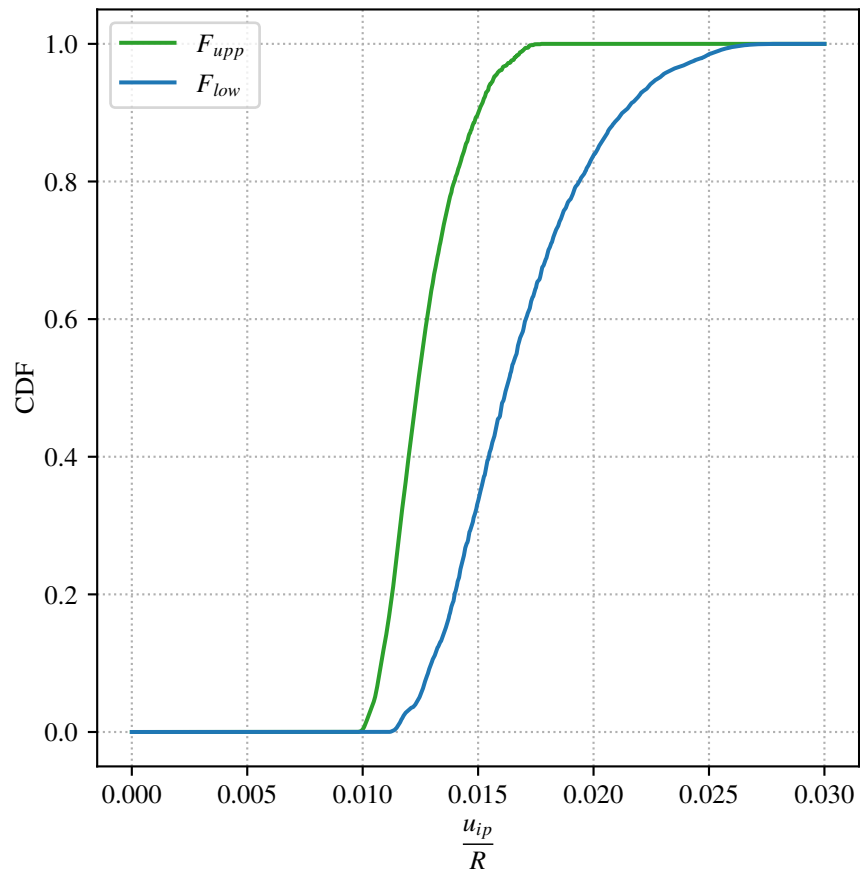
According to Tonon (2004), it is more efficient to increase the fineness of the parameter discretizations above explained, e.g. by employing much more than  $n = 10$  focal elements, than to invoke a global optimization tool or the sampling method. Indeed, by increasing the number of focal elements in the parameter discretizations one achieves a better approximation to the assigned structure of information for the parameter, a finer granularity of the CDFs of  $u_{ip}/R$ , and a better approximation to the global minima and maxima of  $u_{ip}/R$ .

By applying the extension principle to map the random relation to the mechanical system response, both aleatory and epistemic uncertainties of the input parameters are propagated through the equation (5-13). Therefore, although the CDF of  $u_{ip}/R$  is not unique, it is known that it is bounded between a lower and an upper CDFs. In this way,  $(\mathcal{R})$  is used in equation (5-4) and equation (5-5) to define these two bounds. Figure 5-11 shows the resulting limit CDFs,  $F_{upp}$  and  $F_{low}$ , obtained from the uncertain input parameters and after following all the procedures described in this section.

Furthermore, it is possible to compute the expected value of  $u_{ip}/R$  based on the obtained results. For this purpose the following equation must be employed (Tonon, 2004):

$$\mu = \left[ \sum_{i=1}^N m_i \inf(A_i), \sum_{i=1}^N m_i \sup(A_i) \right]$$

For our case, it is found that the expectation of  $u_{ip}/R$  is in  $[1.265 \times 10^{-2}, 1.674 \times 10^{-2}]$ . Note that the expectation of  $u_{ip}/R$  is an interval rather than a unique value. This is because there are both aleatory and epistemic uncertainties in the definition of the parameters employed in equation (5-13), and hence there could not be defined a unique value.



**Figure 5-11:** Example 5.3.1: Upper ( $F_{upp}$ ) and lower  $F_{low}$  bounds obtained for the ratio of the inward displacement of the tunnel lining to the tunnel radius, i.e.  $u_{ip}/R$ .

# 6 Proposed approach for rigorous and efficient geotechnical reliability analyses

So far, we have exposed three different methodologies that complement the conventional probabilistic analyses carried out in geotechnical engineering. In the first instance, subset simulation algorithm was introduced, which allows the probability of failure of geotechnical models to be calculated efficiently. On the other hand, copula theory and its applicability in geotechnical engineering were exhaustively exposed, thus demonstrating the importance of dependence in our area of study and its impact on probabilistic models and analyses. Finally, random sets theory was introduced, as a method that includes both aleatory and epistemic uncertainty in the calculations of probabilities of failure. This chapter will show how these three methodologies can be integrated, thus generating an integral procedure with which efficient reliability analyses can be carried out considering both aleatory and epistemic uncertainties, as well as the dependence between the variables of the models. Most of this chapter is based on the postulates of [Alvarez and Hurtado \(2014\)](#); [Alvarez et al. \(2018\)](#) and [Alvarez \(2007\)](#).

## 6.1 Theoretical basis of the procedure

By using random sets theory, it is not possible to compute an exact value of the probability of failure  $P_X(F)$  of a geotechnical model but only its upper and lower bounds, since the exact multivariate CDF of the input variables is not known with total certainty. Recalling equations (5-5) and (5-4), these bounds can be estimated as follows ([Dempster, 1967](#)):

$$\underline{P}_\Gamma(F) := P_\Omega(\{\alpha \in \Omega : \Gamma(\alpha) \subseteq F, \Gamma(\alpha) \neq \emptyset\}) \quad (6-1)$$

$$\overline{P}_\Gamma(F) := P_\Omega(\{\alpha \in \Omega : \Gamma(\alpha) \cap F \neq \emptyset\}) \quad (6-2)$$

where  $\underline{P}_\Gamma(F) \leq P_X(F) \leq \overline{P}_\Gamma(F)$ .

The evaluation of the equations (6-1) and (6-2) is not straightforward, and it is required other representation for this equations. However, this definition can be particularized depending on

the characteristics of the probability space we want to model. By correctly particularizing the probability space, the evaluation of equations (6-1) and (6-2) can be easier.

In this way, and considering that  $\mu_C$  denotes the Lebesgue-Stieltjes measure associated to the copula  $C$ , which generalizes the  $V_C$ -volume to sets in  $[0, 1]^d$  that are not  $d$ -boxes, Alvarez (2006, 2007) proved that a probability space can be particularized as,

- $\Omega := (0, 1]^d$ ,
- $\sigma_\Omega := \mathcal{B}(\Omega)$ , where  $\mathcal{B}(\Omega)$  is the Borel  $\sigma$ -algebra on  $\Omega$ ,
- $P_\Omega := \mu_C$  for some copula function  $C$  which contains the dependence information within the joint random set,
- $d$ -boxes as elements of  $\mathcal{F}$ ,

in order to model the different representations of uncertainty introduced in section 5.2.2, i.e., possibility distributions, probability boxes, families of intervals, and probability functions. Note that  $\mu_C(F)$  is the probability measure of the region  $F \in [0, 1]^d$  under the copula  $C$  (see Alvarez, 2009, for further details).

This particularization allows to introduce a new representation of the space of the basic variables ( $\mathcal{X}$ -space), which will be called from here on the  $\Omega$ -space. In the  $\Omega$ -space, every focal element from the  $\mathcal{X}$ -space can be represented as a single point, and hence, the computation of equations (6-1) and (6-2) can be simplified. Furthermore, the  $\Omega$ -space allows a more direct integration with the dependence of the basic variables than the  $\mathcal{X}$ -space. In fact, the  $\Omega$ -space contains all the dependence information within the joint random set.

The  $\Omega$ -space plays a key role for integrating random sets theory, copula theory, and subset simulation, in a robust methodology for estimating reliability of geotechnical models. Next, the previous concepts will be deepened and their usefulness in the reliability analyses carried out in geotechnics will be explained.

### 6.1.1 Indexation and sampling from a random sets

Let us consider an universal non-empty set  $X \subseteq \mathbb{R}$ , its power set  $\mathcal{P}(X)$ , and the particularization of the probability space given by Alvarez (2006, 2007), i.e.,  $\Omega := (0, 1]^d$ ,  $\sigma_\Omega := \mathcal{B}(\Omega)$ , and  $P_\Omega := \mu_C$ . Under this particularization, to every focal element of a random set defined in the  $\mathcal{X}$ -space, it can be associated an unique number  $\alpha \in (0, 1]$  that exclusively represents that focal element, and that induces an indexing in the random set.

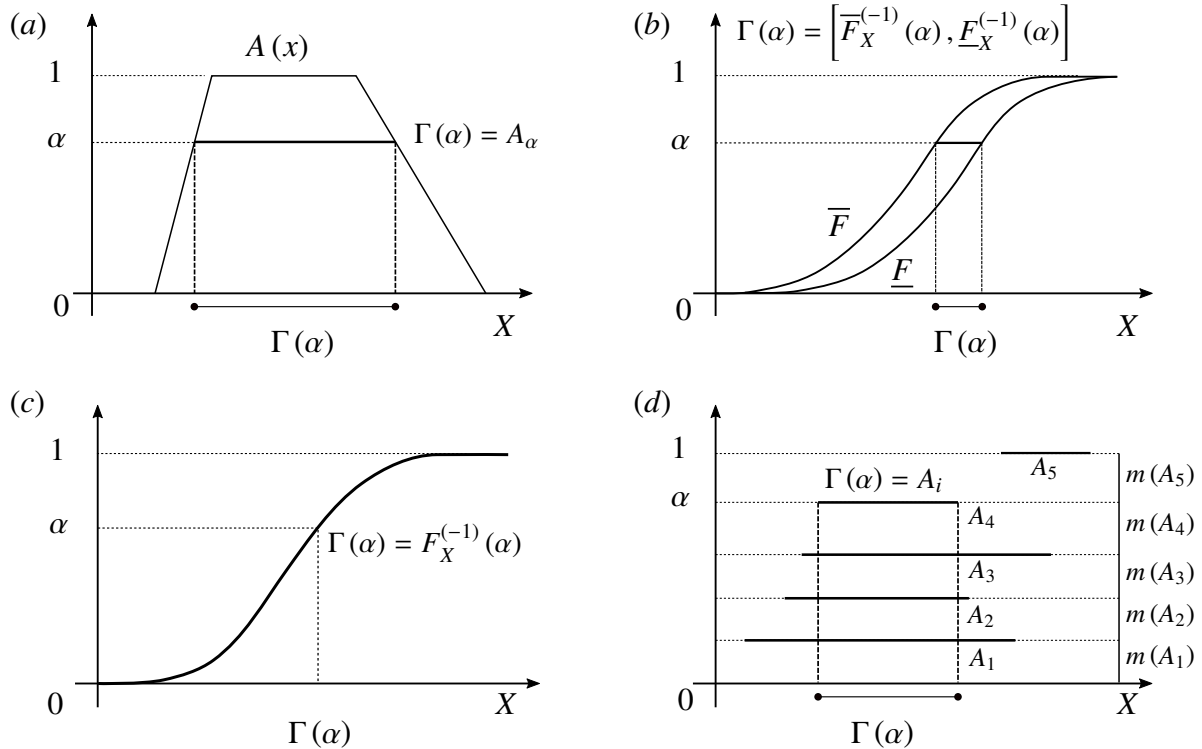
With the previous particularization, it could be introduced a new concept in the construction and handling of random sets, which Alvarez (2007) called *infinite random sets*. An infinite random set is an approximation to a random set through the indexation given by  $\alpha \in (0, 1]$ . The greater the

number of  $\alpha$  sampled, and associated with focal elements in the  $\mathcal{X}$ -space, the better the approximation to the random set  $(\mathcal{F}, P_{\Gamma})$ .

Thus, the lower and upper probability measures of a finite sample  $(\mathcal{F}_n, m)$  will converge almost surely to the lower and upper probability measures of the random set  $(\mathcal{F}, P_{\Gamma})$  as  $N \rightarrow \infty$ . Unlike a finite random set, in which the discretization tends to alter or generate a loss of the information, an infinite random set with an adequate amount of samples overcomes the drawback of loss of information.

In this subsection, we will describe this indexing for four representation of uncertainty, namely possibility distributions, probability boxes, cumulative density functions, and families of intervals. Furthermore, it will also be explained how focal elements can be drawn from each one of these representations of uncertainty through a direct relation with the number  $\alpha \in (0, 1]$ .

In the following,  $\hat{\alpha}$  will denote the random variable, meanwhile  $\alpha$  will denote a realization of that random variable. Furthermore,  $\mathcal{U}_{\hat{\alpha}}(\alpha) := P_{\Omega}(\hat{\alpha} \leq \alpha)$ , with  $\alpha \in (0, 1]$ , defines a CDF on  $(0, 1]$ . All the proof of the concept of this section are contained in (Alvarez, 2007).



**Figure 6-1:** Sampling of focal elements. (a) from a possibility distribution, (b) from a probability box, (c) from a CDF, and (d) from a family of intervals. (adapted from Alvarez, 2007)

### 6.1.1.1 Indexation and sampling from a possibility distribution

Let  $A$  be a possibility distribution with a membership function  $A(x)$  on  $X \in \mathbb{R}$ , and  $\mathcal{F}$  be the family of all  $\alpha$ -cuts  $A_\alpha$ , i.e.,  $\mathcal{F} := \{\Gamma(\alpha) := A_\alpha : \alpha \in (0, 1]\}$ . Then, the possibility distribution can be represented by a RS  $(\mathcal{F}, P_\Gamma)$ , where  $\Gamma(\alpha) = A_\alpha$  and  $P_\Gamma : \sigma_{\mathcal{F}} \rightarrow [0, 1]$  is obtained through equation (5-3).

Thus, for every  $\alpha$  drawn from  $\mathcal{U}_{\hat{\alpha}}$ , there is a unique  $\alpha$ -cut, and viceversa. In this way, there is a one to one relationship between  $A_\alpha$  and  $\alpha$ . Note that  $\alpha$  induces an ordering in  $\mathcal{F}$ , such that  $\alpha_i \leq \alpha_j$  if  $A_i \supseteq A_j$ . Further details of this indexation can be found in Alvarez (2007).

Considering this indexation, sampling a focal element from a possibility distribution is straightforward. Basically, a focal element from a possibility distribution is obtained by drawing a realization of  $\hat{\alpha}$  from  $\mathcal{U}_{\hat{\alpha}}$  and relating the corresponding  $\alpha$ -cut  $A_\alpha$ . Note that as the number of focal elements drawn approaches to infinity ( $N \rightarrow \infty$ ), then the lower and upper probability measure of the finite sample  $(\mathcal{F}_n, m)$  will converge almost surely to the necessity and possibility of the set  $F$  with respect to the possibility distribution  $A$ . An illustration of the sampling from a possibility distribution set is presented in figure 6-1a.

### 6.1.1.2 Indexation and sampling from a distribution-free probability box

Let us consider the distribution-free probability box  $\langle \underline{F}, \overline{F} \rangle$  defined on a subset of  $\mathbb{R}$ . Based on it, let us define the focal element

$$\mathcal{F} = \left\{ \Gamma(\alpha) := \langle \underline{F}, \overline{F} \rangle^{-1}(\alpha) := \left[ \overline{F}^{(-1)}(\alpha), \underline{F}^{(-1)}(\alpha) \right] : \alpha \in (0, 1] \right\}$$

where  $\underline{F}^{(-1)}(\alpha)$  and  $\overline{F}^{(-1)}(\alpha)$  are the quasi-inverses of  $\underline{F}$  and  $\overline{F}$ , respectively. In this way, the probability box  $\langle \underline{F}, \overline{F} \rangle$  can be represented by a random set  $(\mathcal{F}, P_\Gamma)$ , such that  $\Gamma(\alpha) = \langle \underline{F}, \overline{F} \rangle^{-1}(\alpha)$  and  $P_\Gamma$  is obtained through equation (5-3).

Thus, for every  $\alpha$  randomly drawn from  $\mathcal{U}_\alpha$ , there is a unique focal element  $\langle \underline{F}, \overline{F} \rangle^{-1}(\alpha)$ , and this relationship is one to one if the subindex  $\langle \cdot, \cdot \rangle^{-1}(\alpha)$  is preserved.

Futhermore, note that in this particular case  $\alpha$  induces a partial ordering in  $\mathcal{F}$  in the sense that if  $[a_1, b_1]_{\alpha_1}$  and  $[a_2, b_2]_{\alpha_2}$  belong to  $\mathcal{F}$ , then it follows that if  $a_1 < a_2$  then  $a_1 \leq a_2$  and  $b_1 \leq b_2$ . Further details of this indexation can be found in Alvarez (2007).

Considering this indexation, sampling a focal element from a probability box is straightforward. Basically, a focal element from a probability box is obtained by drawing a realization of  $\hat{\alpha}$  from  $\mathcal{U}_{\hat{\alpha}}$  and then retrieve the associated interval  $\langle \underline{F}, \overline{F} \rangle^{-1}(\alpha)$ . This interval will contain the samples for all the CDFs contained in the probability box, i.e.,  $\langle \underline{F}, \overline{F} \rangle^{-1}(\alpha) = \{x : F(x) = \alpha, F \in \langle \underline{F}, \overline{F} \rangle\}$ .

Note that as the number of focal elements drawn approaches to infinity, that is  $N \rightarrow \infty$ , the lower and upper measures of  $(-\infty, x]$  from the sampled RS will converge almost surely to  $\underline{F}(\alpha)$  and  $\overline{F}(\alpha)$ , respectively. An illustration of the sampling from a probability box is presented in figure 6-1b.

### 6.1.1.3 Indexation and sampling from a CDF

The probability law of a continuous random variable ( $X \in \mathbb{R}$ ) can be expressed through a cumulative density function, which is given by  $F_X(x) = P_\Gamma(X \leq x)$  such that  $x \in X$  has some probability measure  $P_\Gamma$ . The CDF of the random variable  $X$ , i.e. ( $F_X$ ) will have a quasi-inverse given by  $F_X^{(-1)}$ . Now, considering  $\hat{\alpha}$  as an uniformly distributed random variable in  $(0, 1]$ , then  $X := F_X^{(-1)}(\hat{\alpha})$  will be distributed as  $F_X$ . Analogously, if  $X$  is a random variable with a continuous CDF  $F_X$ , then a realization of that random variable  $F_X(x)$  is uniformly distributed in  $(0, 1]$ . Therefore, if

$$\mathcal{F} = \{x \mid x := F_X^{(-1)}(\alpha), \alpha \in (0, 1]\}$$

then the random variable  $X$  can be associated with a RS  $(\mathcal{F}, P_\Gamma)$ , such that  $\Gamma(\alpha) = F_X^{(-1)}(\alpha)$  and  $P_\Gamma = \alpha$ . Furthermore, note that if  $\alpha$  has an associated  $x_i$ , then  $x_1 \leq x_2$  if  $\alpha_1 \leq \alpha_2$ , for  $i = 1, 2$ .

In this way, sampling a focal element from a CDF is straightforward. Basically, it is just needed to sample  $\alpha$  from an uniform CDF on  $(0, 1]$ , i.e., from  $\mathcal{U}_\alpha$ , and then select the associated focal element from  $\mathcal{F}$ . Note that as the number of focal elements drawn  $N$  approaches to infinity, the sampled RS will converge almost surely to  $F(X)$ . An illustration of the sampling from a CDF is presented in figure 6-1c.

### 6.1.1.4 Indexation and sampling from a finite family of intervals

In order to index a finite family of intervals  $(\mathcal{F}, m')$  by  $\alpha$ , it is necessary to induce a reproducible and unique ordering. For this purpose, if  $\{[a_i, b_i], \text{ for } i = 1, \dots, s\}$  are the enumeration of the focal elements from  $\mathcal{F}$ , then the family of intervals can be sorted by the criteria:  $[a_i, b_i] \leq [a_j, b_j]$  if  $a_i \leq a_j$  or  $b_i \leq b_j$  when  $a_i \leq a_j$ . This sorting can be performed using a standard sorting procedure (e.g., the quicksort algorithm) by employing an adequate comparison function. More details can be found in Alvarez (2007).

In this way, a finite random set  $(\mathcal{F}_s, m')$  can be represented by  $\Gamma(\alpha) := A^*(\alpha)$ , for  $\alpha \in (0, 1]$ . Note that  $A^*(\alpha)$  is the focal element for which  $\sum_{k=i_1}^{i_1-1} m'(A_k) < \alpha \leq \sum_{k=1}^{i_1} m'(A_k)$ , considering that  $k$  takes successively values from the list  $i_1, i_2, \dots, i_s$ . Note that by convention  $\sum_{k=i_1}^0 m'(A_k) = 0$ .

Now, a focal element of  $\mathcal{F}$  can be obtained (in order to form the RS  $(\mathcal{F}, m)$ ) by drawing an  $\alpha$  from a uniformly distributed CDF on  $(0, 1]$ , i.e.,  $\mathcal{U}_{\hat{\alpha}}$ , and then selecting the respective element  $\Gamma(\alpha)$ . Note that as the number of focal elements drawn approaches to infinity ( $N \rightarrow \infty$ ), then the lower and upper probability measures with respect to the constructed RS  $(\mathcal{F}_n, m)$  will converge almost surely to the belief and plausibility measures of the DS structure  $(\mathcal{F}_s, m')$ . An illustration of the sampling from a finite family of intervals is presented in figure 6-1d.

### 6.1.2 Bounding of the probability of failure using random sets

Under the particularization of the probability space stated by Alvarez (2006, 2007), and considering a failure region in the  $\mathcal{X}$ -space, the sample space  $\Omega$  will contain two subregions given by:

$$F_{LP} := \{\alpha \in \Omega : \Gamma(\alpha) \subseteq F, \Gamma(\alpha) \neq \emptyset\}$$

$$F_{UP} := \{\alpha \in \Omega : \Gamma(\alpha) \cap F \neq \emptyset\}$$

where  $F_{LP}$  and  $F_{UP}$  are formed by all those points in the  $\Omega$ -space that are either completely contained in the set  $F$  or have at least one point in common with the set  $F$  of the  $\mathcal{X}$ -space, respectively. It is worth noting that  $F_{LP} \subseteq F_{UP}$ .

For a bidimensional case, an illustration of the  $\mathcal{X}$ -space and the  $\Omega$ -space, as well as the failure region previously defined, is presented in figure 6-2 a and b, respectively. Note that the realizations of the basic variables in the  $\mathcal{X}$ -space are focal elements depicted as boxes, meanwhile these realizations are points in the  $\Omega$ -space, which are part the copula  $\mathbf{C}$  that defines this space.

In this case, equation (6-1) and (6-2) can be computed using the Lebesgue-Stieltjes integral, as follows (Alvarez, 2006, 2009):

$$\underline{P}_f = \int_{\Omega} \mathbb{I}_{F_{LP}}[\alpha] d\mathbf{C}(\alpha) = \mu_{\mathbf{C}}(F_{LP}) \quad (6-3)$$

$$\overline{P}_f = \int_{\Omega} \mathbb{I}_{F_{UP}}[\alpha] d\mathbf{C}(\alpha) = \mu_{\mathbf{C}}(F_{UP}) \quad (6-4)$$

considering that  $F_{LP}$  and  $F_{UP}$  are  $\mu_{\mathbf{C}}$ -measurable sets.

In this way, it is possible to estimate the lower and upper probabilities of failure using equations (6-3) and (6-4), as follows:

$$\underline{P}_f = \underline{P}_{\Gamma}(F), \quad \overline{P}_f = \overline{P}_{\Gamma}(F) \quad (6-5)$$

Furthermore, it can be seen that the following relations hold, as stated by Alvarez et al. (2018):

$$\mathbb{I}[\Gamma(\alpha) \subseteq F] = \mathbb{I}_{F_{LP}}[\alpha] = \mathbb{I}[\overline{G}(\alpha) \leq 0]$$

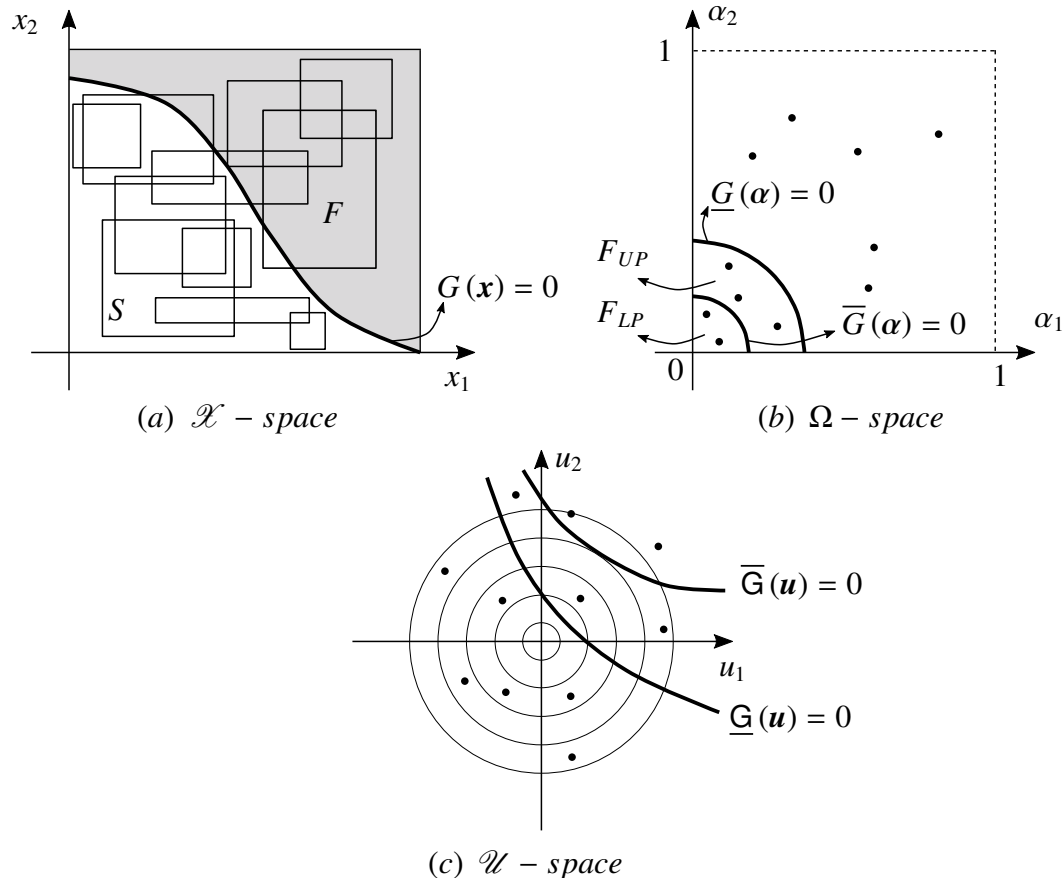
$$\mathbb{I}[\Gamma(\alpha) \cup F \neq \emptyset] = \mathbb{I}_{F_{UP}}[\alpha] = \mathbb{I}[\underline{G}(\alpha) \leq 0]$$

Therefore, as stated by Alvarez and Hurtado (2014), the bounds of the probability of an event of failure (equations (6-5)) can be expressed as:

$$\underline{P}_f = \int_{\Omega} \mathbb{I}[\overline{G}(\alpha) \leq 0] d\mathbf{C}(\alpha) \quad (6-6)$$

$$\overline{P}_f = \int_{\Omega} \mathbb{I}[\underline{G}(\alpha) \leq 0] d\mathbf{C}(\alpha) \quad (6-7)$$

Equations (6-6) and (6-7) state that the computation of  $\underline{P}_f$  and  $\overline{P}_f$  can be expressed through the estimation of two different limit state functions in the  $\Omega$ -space, namely  $\overline{G}$  and  $\underline{G}$ . This fact is a great advantage since it allows us to employ the already existing methods for the computation of probabilities of failure (see section 2.2.2) with some minor modifications.



**Figure 6-2:** The three spaces used to solve the interval reliability problem. (a)  $\mathcal{X}$ -space, or space of the basic variables, in which the focal elements are depicted as boxes (b)  $\Omega$ -space, in which the dependence of the basic variables is defined by the copula function  $C$ , and where each point depicts a focal element, and (c)  $\mathcal{U}$ -space, or the standard Gaussian probability space, where the subsim algorithm is carried out. (adapted from Alvarez and Hurtado, 2014)

### 6.1.3 Monte Carlo simulation for the estimation of $\underline{P}_f$ and $\overline{P}_f$

In this section, we will explain how Monte Carlo simulation can be employed for estimating  $\underline{P}_f$  and  $\overline{P}_f$  through equations (6-6) and (6-7). In fact, the principle of the Monte Carlo simulation does not change, and the minor discrepancies with the conventional method are due the particularization

of the probability space given by  $\Omega := (0, 1]^d$ ,  $\sigma_\Omega := \mathcal{B}(\Omega)$ , and  $P_\Omega := \mu_C$ .

In this regard, Alvarez (2006, 2007, 2009) showed that the equations (6-6) and (6-7) can be computed using the consistent and unbiased estimators:

$$\hat{P}_f = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{F_{LP}}[\alpha_i] = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\overline{G}(\alpha_i) \leq 0] \quad (6-8)$$

$$\overline{P}_f = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{F_{UP}}[\alpha_i] = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[\underline{G}(\alpha_i) \leq 0] \quad (6-9)$$

where  $\alpha_i$  is a sample from the  $d$ -dimensional copula  $C$ .

Note that  $\alpha$  is obtained by sampling it from the copula function  $C$  that defines the  $\Omega$ -space. There are multiple algorithms in the literature to obtain samples of a given copula (see, e.g., Nelsen, 2007; Embrechts et al., 2001). Furthermore, it is worth mentioning that every component  $\alpha_j$ , with  $j = 1, \dots, d$ , from the vector  $\alpha_i$ , is associated with a focal element in the  $\mathcal{X}$ -space, as explained in section 6.1.1. The combination of these focal elements conforms a joint focal element, which is given by the cartesian product  $\times_{j=1}^d \Gamma_j(\alpha_j)$ , where  $\Gamma_j(\alpha_j)$  are the sampled focal elements from each basic variable. This relationship is schematized for a two dimensional case in figure 6-2 a and b. It is worth mentioning that all possible realizations of the input variables for a given  $\alpha_i$  are contained in this joint focal element.

Now, according to equations (6-8) and (6-9),  $\overline{G}(\alpha) > 0$  is an equivalent expression to  $\Gamma(\alpha) \not\subseteq F$ , and also  $\underline{G}(\alpha) > 0$  and  $\Gamma(\alpha) \cap F = \emptyset$  are equivalent expressions. Additionally, note that if  $\Gamma(\alpha) \cap F = \emptyset$ , it implies that  $\Gamma(\alpha) \not\subseteq F$ , and hence,  $\underline{G}(\alpha) > 0$  entails  $\overline{G}(\alpha) > 0$ . In this way, the number of evaluations of the function  $\overline{G}$  can be significantly reduced by considering this principle. In other words, when a sample  $\alpha_i$  from the copula  $C$  is employed in the evaluation of the equations (6-8) and (6-9), and it satisfies  $\mathbb{I}[\underline{G}(\alpha_i) \leq 0]$ , then it is not necessary to evaluate  $\overline{G}$ . Therefore, equation (6-9) must be evaluated first, and only those samples  $\alpha_i$  that fulfill  $\underline{G}(\alpha_i)$  would have to be employed in the evaluation of  $\overline{G}$  through the equation (6-8).

On the other hand, it is worth emphasizing again that the larger the sample size  $N$  in equations (6-8) and (6-9), the better the approximation to the bounds of the probability of failure, such as in a conventional Monte Carlo simulation. In this way, if the expected bounds of the probability are small (especially in the case of  $\underline{P}_f$ ), more samples would be required in order to guarantee a good accuracy in the estimates.

Finally, note that equations (6-8) and (6-9) can also be computed by any advance Monte Carlo Simulation method, such as importance sampling, line sampling, adaptive sampling, or subset simulation. This latter in particular will be the method we will address below.

## 6.2 Proposed approach: efficient estimation of $\underline{P}_f$ and $\overline{P}_f$ using subset simulation

It was previously explained how the Monte Carlo simulation can be used to estimate the bounds of the probabilities of failure through the equations (6-8) and (6-9), thus integrating random sets theory with copula theory. However, as discussed in section 3, this simulation can be quite expensive, computationally speaking, and it is inefficient when the probability of failure is small. Therefore, more efficient but precise methods should be used, such as the subset simulation algorithm, which was introduced in section 3.

To integrate the subset simulation algorithm with random sets theory and copula theory, in order to generate a robust but efficient method for reliability analysis in geotechnics, it is necessary to carry out the algorithm in a new probability space. For this purpose, a wise option is the standard Gaussian space  $\mathcal{U} \in \mathbb{R}^d$ , which is widely used by many methods that seek to estimate probabilities of failure (Alvarez et al., 2018).

In this way, it is required to perform a transformation of variables between the  $\Omega$ -space and the  $\mathcal{U}$ -space. This isoprobabilistic transformation can be carried out by means of a bijective function  $T$  that maps the point  $\alpha \in \Omega$ , defined by the copula  $C$ , to an independent point in  $(0, 1]^d$ , as explained by Lemaire (2013, chapter 4) or Ditlevsen and Madsen (1996, chapter 7). So, by defining  $\alpha = T^{-1}(\Phi(\mathbf{u}))$ , equations (6-6) and (6-7) become

$$\underline{P}_f = \int_{\mathcal{U}} \mathbb{I}[\underline{\mathbf{G}}(\mathbf{u}) \leq 0] d\Phi_U(\mathbf{u}) \quad (6-10)$$

$$\overline{P}_f = \int_{\mathcal{U}} \mathbb{I}[\overline{\mathbf{G}}(\mathbf{u}) \leq 0] d\Phi_U(\mathbf{u}) \quad (6-11)$$

where the limit state functions,  $\underline{\mathbf{G}}$  and  $\overline{\mathbf{G}}$ , are given by

$$\begin{aligned} \underline{\mathbf{G}}(\mathbf{u}) &= \underline{G}(T^{-1}(\Phi(\mathbf{u}))) \\ \overline{\mathbf{G}}(\mathbf{u}) &= \overline{G}(T^{-1}(\Phi(\mathbf{u}))) \end{aligned}$$

such that  $\Phi_U$  represents the standard Gaussian joint CDF,  $\Phi(\mathbf{u}) = [\Phi(u_1), \dots, \Phi(u_d)]$ , and  $\Phi$  is the one-dimensional standard Gaussian CDF.

These two limit state functions are schematized, for a bidimensional case, in figure 6-2c. In this figure, the concentric circles represent the contours of the bivariate standard Gaussian distribution. Furthermore, it is worth mentioning again that each point in the  $\mathcal{U}$ -space (figure 6-2c) is related to a point in the  $\Omega$ -space (figure 6-2b), and these in turn are related to a joint focal element in the  $\mathcal{X}$ -space (figure 6-2a).

In this way, the standard Gaussian variables in the  $\mathcal{U}$ -space are transformed to the independent space  $[0, 1]^d$  through the operation  $\Phi(\mathbf{u})$ . Afterwards, the inverse transformation  $T^{-1}$  is employed

to map each independent point in  $[0, 1]^d$  to a dependent one in the  $\Omega$ -space, according to the copula  $C$ .

Finally, it is worth mentioning that there are many algorithms in the literature for transforming a vector of independent points  $[0, 1]^d$  into a vector of points that follow the structure of dependence of a copula function. For instance, one option would be to perform the transformation of  $T$  using the Rosenblatt transformation  $T_{Ross}$ , as explained by [Lemaire \(2013, chapter 4\)](#). Readers are referred to [Nelsen \(2007\)](#) or [Embrechts et al. \(2001\)](#) for an introduction to these algorithms.

## 6.3 Numerical examples

The applicability of the approach proposed in this thesis will be illustrated in this section through three examples of geotechnical models, quite common in practice. These three examples integrate both aleatory and epistemic uncertainties (through random sets theory), take into account the dependence between the variables (through copula theory), and the simulation of them is accurate and efficient thanks to the subset simulation algorithm.

For the subset simulation algorithm, a conditional probability of  $p_0 = 0.1$  is employed for defining each intermediate level, in each of which the image of  $N = 1000$  focal elements are computed. Furthermore, a uniform proposal distribution with width 2 is employed in the algorithm.

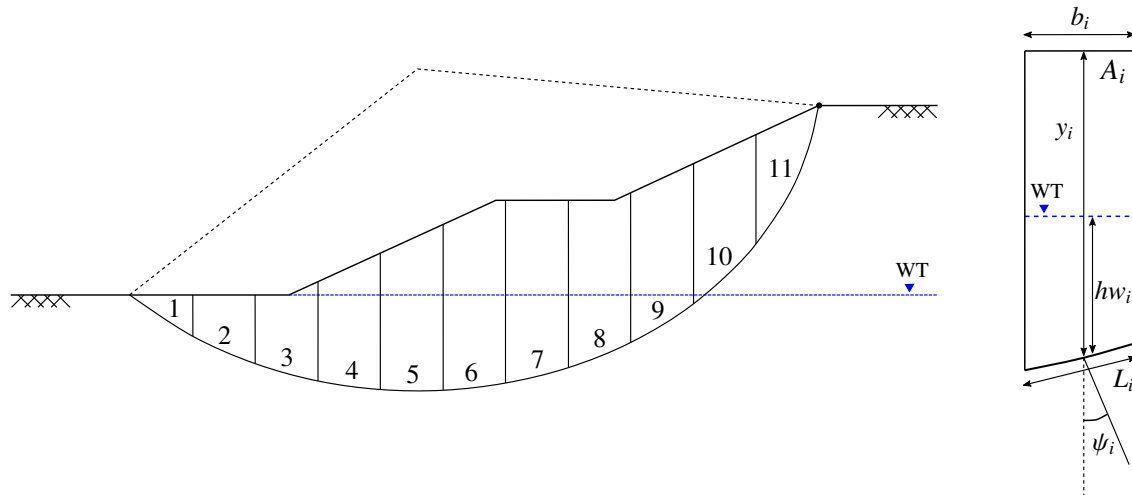
Regarding random sets case, it is necessary to use of the extension principle in order to obtain the lower and upper LSFs. In the examples of this section, the image of each focal element  $\Gamma(\alpha)$  through the LSF  $G$  is found employing the sampling method explained in section 5.2.6. For this method, a sample of 100 elements is drawn from the domain of the focal element using a uniform probability distribution. This method, and this sample size, are used in the following examples since better precision is achieved than in the vertices method, but at the same time the computational cost is lower than the optimization method.

### 6.3.1 Example 1: Rotational failure of an homogeneous soil slope

Let us consider the finite homogeneous slope of figure 6-3, for which a rotational failure mechanism was established. This slope is constituted by a soil that follows a Mohr-Coulomb failure criterion, that is to say, it is characterized by a cohesion  $c$ , an inner friction angle  $\phi$ , and a unit weight  $\gamma$ . However, although it is known the unit weight is a deterministic parameter equal to  $20 \text{ kN/m}^3$ , cohesion and inner friction of the soil are uncertain parameters.

Nonetheless, it is defined that  $c$  and  $\phi$  can be modeled as the distributional p-boxes  $\mathcal{N}([15, 20], 3)$  and  $\mathcal{N}([22, 26], 4)$ , respectively, where  $\mathcal{N}$  stands for the Normal distribution. Furthermore, it is also known that the Kendall's tau correlation coefficient between these two variables is equal to  $-0.5$ , and that the structure of dependence of them is given by a Nelsen No. 16 copula.





**Figure 6-4:** Example 1 of the proposed approach: details of the slices employed for the Fellenius method applied in the slope of figure 6-3

For this example, a total number of 11 slices are employed, each one with a width  $b_i = 29/11$  m. These slices are shown in figure 6-4, where a detail of their geometry is also presented. Furthermore, the deterministic properties of each one of these slices are summarized in table 6-1.

On the other hand, note that the distributional p-boxes that describe the parameters  $c$  and  $\phi$  must be converted to distributional-free p-boxes, in order to use them in the simulation. For this purpose, the method described in Alvarez et al. (2017) can be employed.

**Table 6-1:** Example 1 of the proposed approach: properties of each one of the 11 slices of the homogeneous soil slope presented in figure 6-4

Slice	Area (m)	$L_s$ (m)	$h_w$ (m)	$\psi$ ( $^\circ$ )	$W$ (kN/m)	$u_w$ (kN/m $^2$ )	$U$ (kN/m)
1	3.08	3.43	1.16	$-39.65^\circ$	61.60	11.38	39.03
2	7.81	3.01	2.97	$-28.80^\circ$	156.20	29.14	87.70
3	11.31	2.79	4.14	$-19.01^\circ$	226.20	40.61	113.31
4	16.50	2.68	4.82	$-9.80^\circ$	330.00	47.28	126.72
5	21.14	2.64	5.1	$-0.84^\circ$	422.80	50.03	132.08
6	26.64	2.67	4.9	$8.11^\circ$	532.80	48.07	128.34
7	24.50	2.76	4.31	$17.26^\circ$	490.00	42.28	116.70
8	21.81	2.96	3.23	$26.90^\circ$	436.20	31.69	93.79
9	20.38	3.33	1.56	$37.50^\circ$	407.60	15.30	50.96
10	17.76	4.12	0.00	$50.04^\circ$	355.20	0.00	0.00
11	10.00	7.86	0.00	$70.23^\circ$	200.00	0.00	0.00

With these considerations, it is possible to compute equations (6-10) and (6-11) through the proposed methodology. Following the proposed procedure, it is obtained that the estimates of the probability bounds are  $\left[\underline{P}_f, \overline{P}_f\right] = \left[6.7 \times 10^{-3}, 4.38 \times 10^{-4}\right]$ . That is to say, considering epistemic and aleatory uncertainties in the definition of the soil shear strength parameters, and also their dependence, the probability of failure of the homogeneous slope depicted in figure (6-3) is bounded by the above two limits.

Additionally, taking advantage of the fact that this is a two-dimensional model, it is interesting to illustrate how the samples are distributed in each of the three workspaces. That is to say, the idea is to analyze how the samples are distributed in: *a.*) the original space of the basic variables ( $\mathcal{X}$ -space), where the focal elements of the random sets will be boxes in 2 dimensions (rectangles), *b.*) the copula domain ( $\Omega$ -space), which contains the structure of dependence between the basic variables and where each focal elements is specified as a point  $\alpha_i$ , and *c.*) the standard Gaussian domain ( $\mathcal{U}$ -space), in which the reliability problem is solved through the subset simulation algorithm.

Figure 6-5 shows the distribution and propagation of the SubSim samples in each one of these three spaces. Furthermore, the failure surface given by  $G(c, \phi)$  from equation (6-12) is also presented in the  $\mathcal{X}$ -space of this figure.

Note from figure 6-5 that the computation of the upper probability of failure ( $\overline{P}_f$ ) employed three intermediate failure levels, meanwhile the computation of the lower probability of failure ( $\underline{P}_f$ ) employed four intermediate failure regions. From this figure it is also evident how the subset simulation algorithm redirects the simulation to the failure domain, thus considerably improving the efficiency in the calculation of the bounds of the probability of failure.

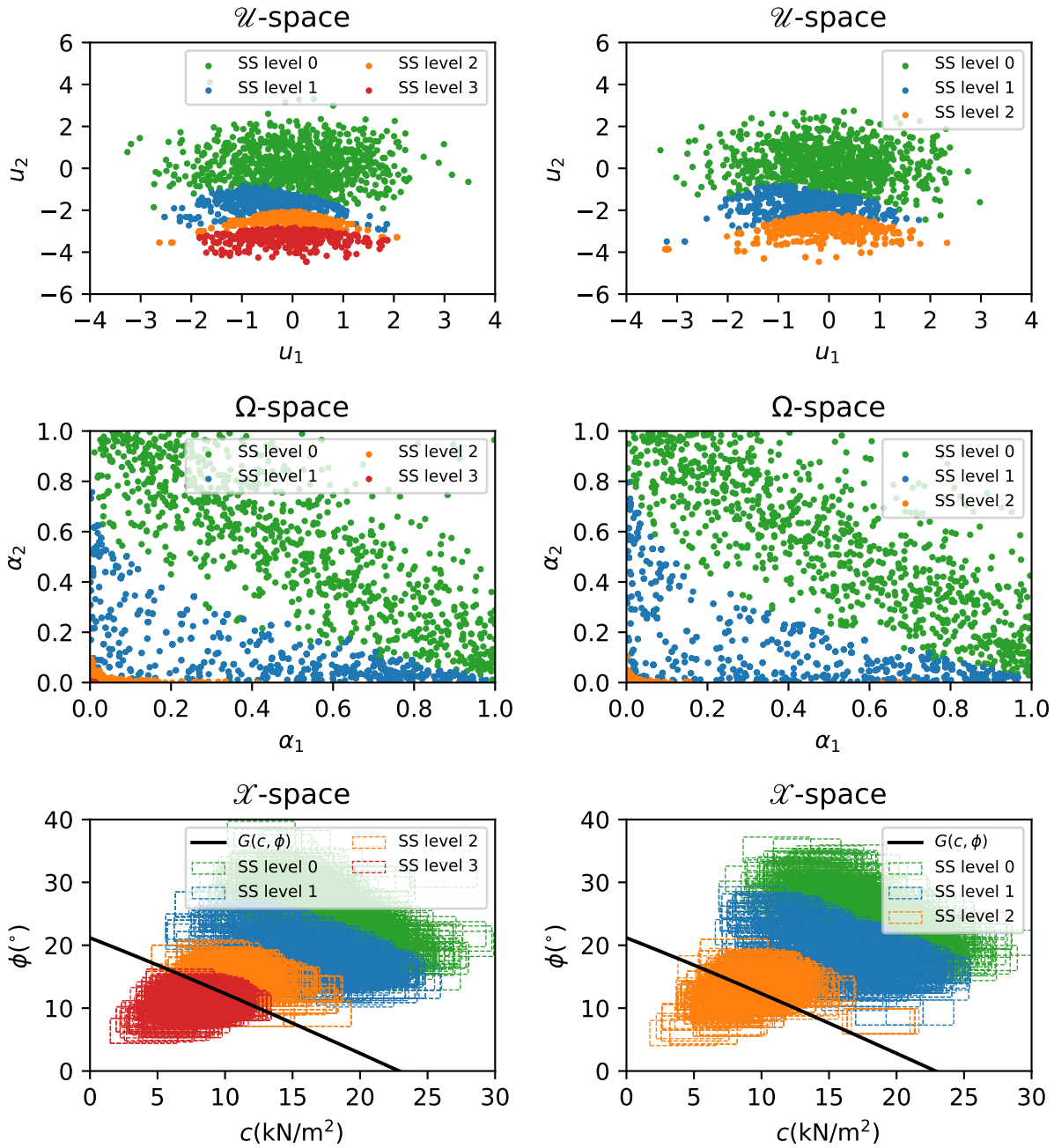
### 6.3.2 Example 2: Toppling failure of a rock slope

It is planned to make a cut in a rocky mass, in such a way that the face of the resulting slope has an inclination of  $\psi_f$  with the horizontal. Given the conditions of the rocky mass, it is identified that the resulting slope has a toppling-type failure mechanism. Figure 6-6 schematizes the rock slope of this example and shows geometric details of its discontinuities and exposed blocks. The numeric values of the geometric parameters from figure 6-6 are presented in table 6-2.

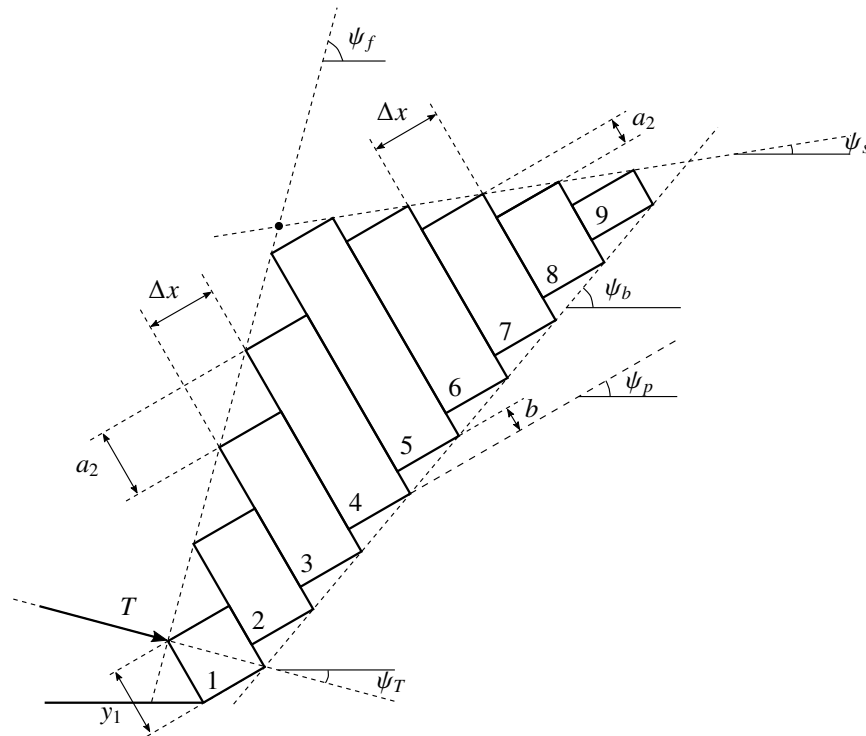
**Table 6-2:** Example 2 of the proposed approach: geometric details of the rock slope presented in figure 6-6

$\psi_f =$	$75^\circ$	$\psi_b =$	$50^\circ$	$\Delta x =$	1.5 m
$\psi_s =$	$9^\circ$	$\psi_p =$	$30^\circ$	$y_1 =$	1.5 m

As an action to stabilize the slope, it is proposed to use horizontal active anchors, with a tension  $T$ , in the first block of the slope (see figure 6-6). It is then necessary to estimate the bounds of the



**Figure 6-5:** Evolution of the SubSim samples in the  $\mathcal{X}$ -space, the  $\Omega$ -space, and the  $\mathcal{U}$ -space, for the computation of the lower bound  $P_f$  (graphs in the left column) and the upper bound  $\overline{P}_f$  (graphs in the right column) of the probability of failure of the model studied in the example 1 of the proposed approach.



**Figure 6-6:** Example 2 of the proposed approach: scheme of a rock slope with a toppling-type failure mechanism.

probability of failure of this model  $(\underline{P}_f, \overline{P}_f)$ , for different values of  $T = [0, 600]$  kN, taking into account the following considerations:

- The friction angle of the rock mass ( $\phi$ ) is given by the distributional p-box  $\mathcal{LN}([25, 30], [3, 5])$ , where  $\mathcal{LN}$  stands for the Lognormal probability distribution.
- The unit weight of the rock mass ( $\gamma$ ) counts with two independent and equally credible sources of information for its definition. Each source defines a closed interval  $\gamma_i$  for the values that  $\gamma$  can adopt. Thus,  $\gamma_1 = [17, 20]$  kN/m<sup>3</sup>, and  $\gamma_2 = [19, 22]$  kN/m<sup>3</sup>.
- An earthquake, characterized by a horizontal coefficient of seismic acceleration  $K_h$ , is expected to impact the slope. The coefficient  $K_h$  can be modeled as a triangular probability distribution  $\mathcal{T}$  with a lower limit of 0.05, an upper limit of 0.20, and a mode of 0.10.
- It is known that  $\phi$  and  $\gamma$  are related by a Gumbel copula adjusted by a Kendall's tau coefficient of 0.75. This leads to the copula  $C_{Gumbel} = (\alpha_1, \alpha_2; 4.0)$ , where  $\alpha_1$  and  $\alpha_2$  represent the variables  $\phi$  and  $\gamma$  in the  $\Omega$ -space, respectively. The horizontal coefficient of seismic acceleration is independent of the rest variables. Therefore, the copula that relates of the implied

variables of the model is:

$$C(\alpha) = C_{Gumbel} = (\alpha_1, \alpha_2; 4.0) \cdot \alpha_3$$

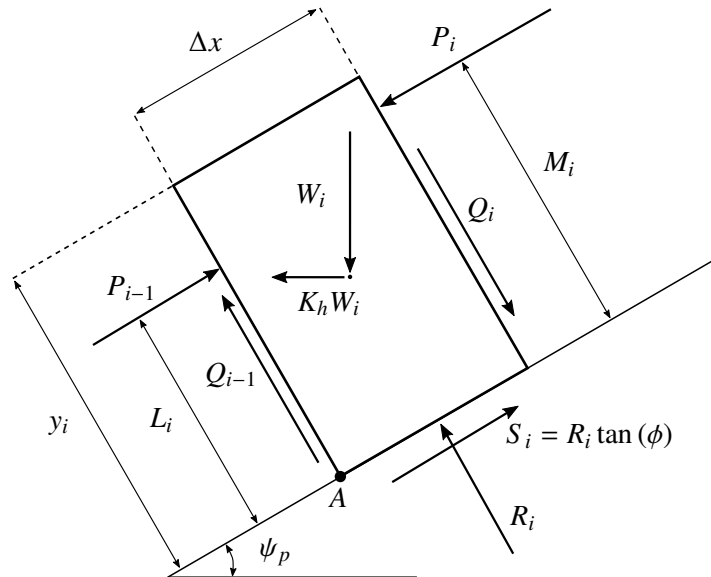
where  $\alpha_1$  and  $\alpha_2$  were previously defined, and  $\alpha_3$  represents the variable  $K_h$  in the  $\Omega$ -space.

In order to conduct the reliability analysis of this problem it is required to define the limit state function (LSF) of the model, which can be expressed as:

$$G(\phi, \gamma, K_h) = FS - 1 \quad (6-14)$$

where  $FS$  is obtained following the procedure described by [Hoek and Bray \(1981, chapter 10\)](#) or [Wyllie \(2017, chapter 10\)](#).

This procedure is based on a limit equilibrium analysis for each of the blocks that are part of the slope, starting with the highest block and ending with the lowest block. [Figure 6-7](#) presents a typical block ( $i$ ) and the forces involved in its limit equilibrium analysis.



**Figure 6-7:** Example 2 of the proposed approach: limit equilibrium analysis of toppling on a steeped base, forces acting in the  $i$ th block.

In [figure 6-7](#),  $Q_i$  and  $Q_{i-1}$  are the shear forces transmitted by the adjacent blocks,  $P_i$  and  $P_{i-1}$  are the normal forces transmitted by the adjacent blocks, and  $R_i$  and  $S_i$  are the normal and shear forces developed on the base of the block. The distances  $M_i$  and  $L_i$  are the points of application of the normal forces  $P_i$  and  $P_{i-1}$ , respectively,  $y_i$  is the block height and  $\Delta x$  is the block width.

The numerical values of  $y_i$ ,  $M_i$  and  $L_i$ , for each block, can be found by simple trigonometry or following the procedure and equations presented in [Wyllie \(2017, chapter 10\)](#), it is also necessary

to consider the following values (see figure 6-6):

$$\begin{aligned} a_1 &= \Delta x \tan(\psi_f - \psi_p) = 1.50 \text{ m} \\ a_2 &= \Delta x \tan(\psi_p - \psi_s) = 0.58 \text{ m} \\ b &= \Delta x \tan(\psi_b - \psi_p) = 0.55 \text{ m} \end{aligned}$$

for an irregular array of blocks,  $y_i$ ,  $M_i$  and  $L_i$  can be found graphically.

Table 6-3 shows the values of  $y_1$ ,  $M_i$  and  $L_i$ , as well as the area, of each block of the example studied in this section.

**Table 6-3:** Example 2 of the proposed approach: heights, areas and point of application of forces for each block in the figure 6-6.

Block	$y_i$ (m)	$M_i$ (m)	$L_i$ (m)	Area (m <sup>2</sup> )
1	1.5	1.5	1.5	2.25
2	2.45	2.45	0.95	3.68
3	3.41	3.41	1.91	5.12
4	4.36	4.36	2.86	6.54
5	5.32	4.74	3.82	7.98
6	4.19	3.62	4.19	6.29
7	3.07	2.5	3.07	4.61
8	1.95	1.38	1.95	2.93
9	0.83	0.25	0.83	1.25

Now, for each block of the rock slope two failure mechanism are studied: sliding and toppling. Sliding is studied taking the sum of forces in the direction of  $S_i$ , and toppling is studied taking the sum of moments in the point A (see figure 6-7). In this way, the force  $P_{i-1,s}$  to guarantee stability against sliding and the force  $P_{i-1,t}$  to guarantee stability against toppling are computed for each block. The equations of these two analyses are (see, e.g., Hoek and Bray, 1981; Wyllie, 2017):

$$P_{i-1,s} = P_i + \frac{K_h W_i (\sin(\psi_p) \tan(\phi) + \cos(\psi_p)) - W_i (\cos(\psi_p) \tan(\phi) - \sin(\psi_p))}{1 - \tan^2(\phi)} \quad (6-15)$$

$$P_{i-1,t} = \frac{P_i (M_i - \Delta x \tan(\phi)) + \left(\frac{W_i}{2}\right) (y_i \sin(\psi_p) - \Delta x \cos(\psi)) + \left(K_h \frac{W_i}{2}\right) (y_i \cos(\psi_p) + \Delta x \sin(\psi_p))}{L_i} \quad (6-16)$$

The greater force between  $P_{i-1,s}$  and  $P_{i-1,t}$  will be the force  $P_i$  of the next block to analyze. This procedure is performed starting in the highest block (in this case the block 9) and ending in the

lowest block (in this case the block 1), finding thus the force  $P_0$  for this latter block. Note that for the block 1 the force transmitted by the anchor must be taken into account in the equations (6-15) and (6-16).

Finally, if  $P_0 < 0.0$  then the slope is stable, otherwise the slope is unstable. Now, following this procedure, the expression of  $FS$  from equation (6-14) can be written as follows:

$$FS = \frac{\tan(\phi)}{\tan(\phi_{required})}$$

where  $\phi$  is the *available friction angle*, i.e. the friction angle of the rock mass, and  $\phi_{required}$  is the friction angle that leads to a  $P_0 = 0.00$ , i.e. the friction angle that generates a limit equilibrium in the system. Note that  $\phi_{required}$  must be found iteratively or using numerical methods.

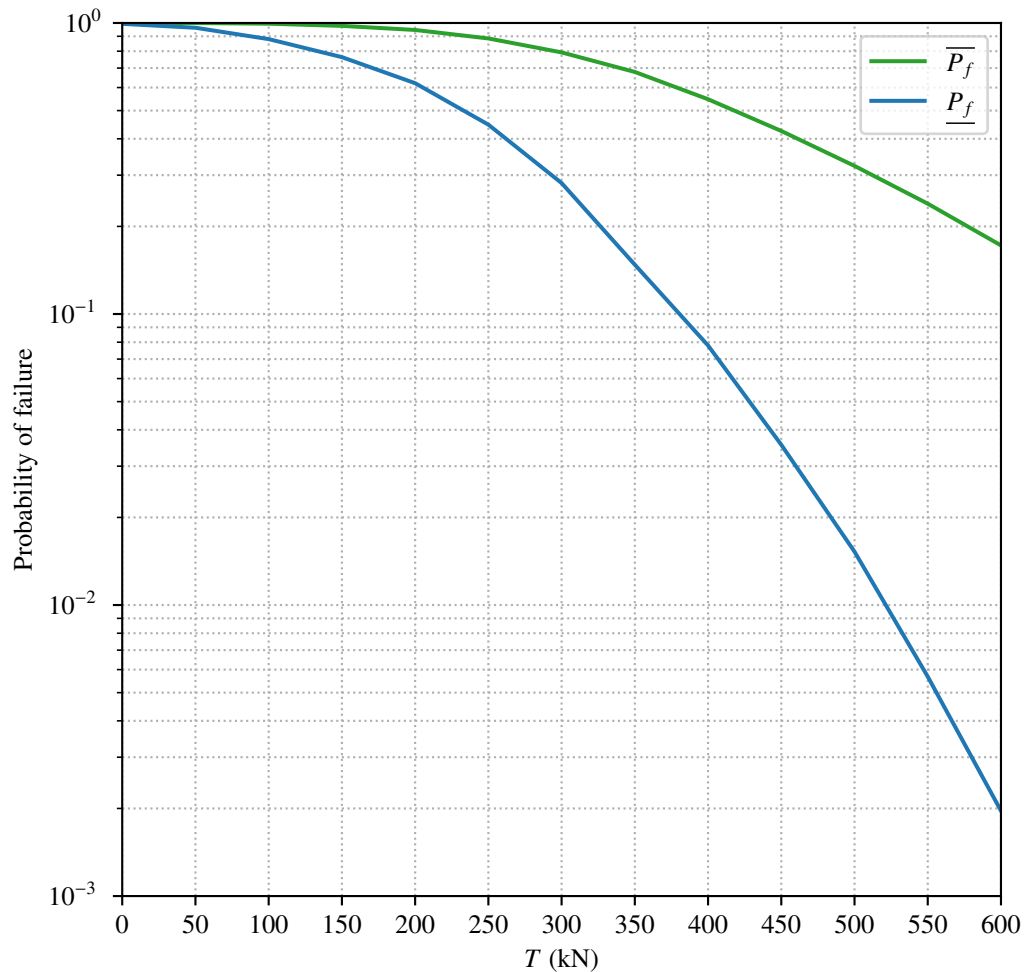
Now, by applying the proposed approach considering this procedure and the conditions already established for the model, it is possible to compute the bounds of the probability of failure of the system, for different loads  $T \in [0, 600]$  kN in the anchor of the first block. Figure 6-8 presents the evolution of these bounds as the magnitude of  $T$  increases. Furthermore, table 6-4 shows the numerical values of the bounds on the probability of failure of the system, for values of  $T$  within  $[0, 600]$  kN in intervals of 50 kN.

**Table 6-4:** Example 2 of the proposed approach: results of the bounds on the probability of failure in the rock slope from figure 6-6, considering values of  $T$  within  $[0, 600]$  kN in intervals of 50 kN.

$T$ (kN)	$\underline{P}_f$	$\overline{P}_f$	$T$ (kN)	$\underline{P}_f$	$\overline{P}_f$
0	$9.932 \times 10^{-1}$	$1 \times 10^0$	350	$1.478 \times 10^{-1}$	$6.788 \times 10^{-1}$
50	$9.630 \times 10^{-1}$	$9.998 \times 10^{-1}$	400	$7.796 \times 10^{-2}$	$5.470 \times 10^{-1}$
100	$8.802 \times 10^{-1}$	$9.944 \times 10^{-1}$	450	$3.554 \times 10^{-2}$	$4.256 \times 10^{-1}$
150	$7.632 \times 10^{-1}$	$9.774 \times 10^{-1}$	500	$1.526 \times 10^{-2}$	$3.228 \times 10^{-1}$
200	$6.211 \times 10^{-1}$	$9.464 \times 10^{-1}$	550	$5.676 \times 10^{-3}$	$2.394 \times 10^{-1}$
250	$4.474 \times 10^{-1}$	$8.848 \times 10^{-1}$	600	$1.964 \times 10^{-3}$	$1.721 \times 10^{-1}$
300	$2.818 \times 10^{-1}$	$7.924 \times 10^{-1}$			

By observing figure 6-8 and table 6-4, it is evident that the system requires an anchor for its stability. When there is no anchor in the system, or when the load transmitted by it is low (say  $T < 250$  kN), the rock slope will fail almost surely. As the load transmitted by the anchor increases the probability of failure is reduced; this is to be expected according to the physical sense of the system.

Now, what is really interesting here is how the bounds of the probability of failure evolve as the load on the anchor increases. It can be seen in figure 6-8 that as the load on the anchor increases, the gap between the bounds of  $P_f$  also increases. This means that epistemic uncertainties have a greater impact when the probability of failure is small, that is, when the system is reliable.



**Figure 6-8:** Example 2 of the proposed approach: evolution on the bounds of the probability of failure in the rock slope from figure 6-6, considering different loads  $T \in [0, 600]$  kN transmitted by an anchor in the first block of the slope.

Based on the values of table 6-4, it could be observed that when  $T \leq 350$  kN the bounds on the probability of failure are in the same order of magnitude; nonetheless, for  $T \geq 400$  kN these bounds are in a different order of magnitude, reaching for example a difference of 2 orders of magnitude when  $T = 600$  kN. If the epistemic uncertainty in the definition of parameters were reduced, the gap between these bounds would also be reduced, thus converging to a single value when the epistemic uncertainty is zero.

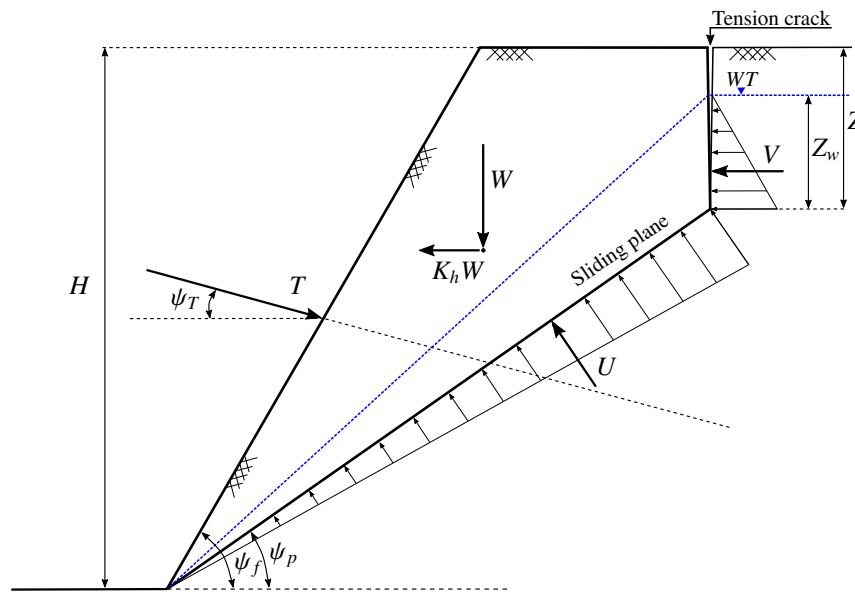
### 6.3.3 Example 3: Plane failure of a rock slope

For the construction of a new highway, it is necessary to carry out an excavation in a rock mass; this procedure will generate a twenty meters high slope ( $H = 20$  m) at a face angle of  $\psi_f = 60^\circ$ . The rock mass where the excavation will take place contains persistent bedding planes that dip at an angle of  $\psi_p = 35^\circ$  into the excavation. It is also known that behind the crest there will be a tension crack with depth  $Z$ , and which will be filled with water to a height of  $Z_w$  above the sliding surface.

It is defined that the rock mass follows a Mohr-Coulomb failure criterion. In other words, the failure of the rock slope can be modelled by a cohesion  $c$ , an inner friction angle  $\phi$ , and a unit weight  $\gamma$ .

Furthermore, it is desired to take into account in the model the impact of an earthquake that would generate a horizontal acceleration in the terrain of  $K_h g$ , where  $K_h$  is the horizontal coefficient of seismic acceleration and  $g$  is the gravity. Taking into account this adverse load, the use of active horizontal anchors is proposed, which transmit a load  $T$  to the wedge, in order to guarantee the stability of the slope.

Figure 6-9 presents an scheme of the rock slope studied in this example. It is worth noting that in this figure the forces  $U$  and  $V$  are the resultant of the distributed forces due to hydrostatic pressures in the sliding plane and the tension crack, respectively. Furthermore, note that  $\psi_T = 0$  since the anchor is said to be horizontal.



**Figure 6-9:** Example 3 of the proposed approach: two-dimensional stability model of a rock slope with a plane failure mechanism.

In a conventional approach, the factor of safety (FS) of the slope can be formulated as follows (see, e.g., [Hoek and Bray, 1981](#); [Wyllie, 2017](#), chapter 7):

$$FS = \frac{cA + N' \tan(\phi)}{W(\sin(\psi_p) + K_h \cos(\psi_p)) + V \cos(\psi_p) - T \cos(\psi_T + \psi_p)} \quad (6-17)$$

where

$$A = \frac{(H - z)}{\sin(\psi_p)}$$

$$N' = W(\cos(\psi_p) - K_h \sin(\psi_p)) - U - V \sin(\psi_p) + T \cos(\psi_T + \psi_f)$$

$$W = 0.5\gamma H^2 \left( \left( 1 - \left( \frac{Z}{H} \right)^2 \right) \cot(\psi_p) - \cot(\psi_f) \right)$$

and

$$U = 0.5\gamma_w r Z A$$

$$V = 0.5\gamma_w r^2 Z^2$$

$$r = \frac{Z_w}{Z}$$

Nonetheless, a deterministic analysis is not recommended in this case since some of the parameters of the model have uncertainties in their definition, whether aleatory uncertainty, epistemic uncertainty, or both. In this regard, it is then required to estimate the bounds of the probability of failure  $(\underline{P}_f, \overline{P}_f)$  of the system, for different values of  $T$  within  $[0, 700]$  kN, considering the following information:

- The unit weight  $\gamma$  of the rock mass that compounds the slope is set as a deterministic value equal to  $26 \text{ kN/m}^3$ .
- The cohesion of the rock mass ( $c$ ) is given by the distributional p-box  $\mathcal{LN}([15, 20], 3) \text{ kN/m}^2$ , where  $\mathcal{LN}$  stands for the Lognormal probability distribution.
- It is defined that the friction angle of the rock mass ( $\phi$ ) is given by the distributional p-box  $\mathcal{N}(35, [4, 6])^\circ$ , where  $\mathcal{N}$  stands for the Normal probability distribution.
- The depth of the tension crack ( $Z$ ) counts with three independent and equally credible sources of information for its definition. Each source sets a closed interval  $Z_i$ , within which the values of  $Z$  are considered to lie. Thus,  $Z_1 = [1, 6] \text{ m}$ ,  $Z_2 = [5, 8] \text{ m}$ , and  $Z_3 = [7, 10] \text{ m}$ .
- The ratio of water depth filled in the tension crack to tension crack depth ( $r$ ) can be modeled as a triangular probability distribution  $\mathcal{T}$  with a lower limit of  $0.00 \text{ m}$ , an upper limit of  $1.00 \text{ m}$ , and a mode of  $0.33 \text{ m}$ . Note that  $r = 0.00$  means that the tension crack is dry, and  $r = 1.00$  means that the tension crack is completely full of water.

- The horizontal coefficient of seismic acceleration is modeled as a uniform probability distribution  $\mathcal{U}$  with a lower limit of 0.05 and an upper limit of 0.15.
- The structure of dependence between  $c$  and  $\phi$  is given by a Frank copula fitted through a Kendall's tau coefficient of  $-0.75$ . This leads to the copula  $C_{Frank}(\alpha_1, \alpha_2; -14.138)$ , where  $\alpha_1$  and  $\alpha_2$  represent the variables  $c$  and  $\phi$  in the  $\Omega$ -space, respectively.
- The structure of dependence between  $Z$  and  $r$  is given by a Plackett copula fitted through a Spearman's rho coefficient of  $-0.5$ . This leads to the copula  $C_{Plackett}(\alpha_3, \alpha_4; 0.195)$ , where  $\alpha_3$  and  $\alpha_4$  represent the variables  $Z$  and  $r$  in the  $\Omega$ -space, respectively. Note that this relationship is negative, which means that a tension crack with a shallow depth is more likely to be completely filled with water, meanwhile a tension crack with a greater depth is more likely to be less full of water.
- It is set that  $K_h$  is independent of the rest of variables of the model. Furthermore, it is known that  $c$  and  $\phi$  are independent of  $Z$  and  $r$ . In consequence, the copula that relates all the implies variables is:

$$C(\alpha) = C_{Frank}(\alpha_1, \alpha_2; -14.138) \cdot C_{Plackett}(\alpha_3, \alpha_4; 0.195) \cdot \alpha_5$$

where  $\alpha_1, \dots, \alpha_4$  were previously defined, and  $\alpha_5$  represents the variable  $K_h$  in the  $\Omega$ -space.

In order to conduct this reliability analysis it is required to define the limit state function of the model, which can be formulated as follows:

$$G(c, \phi, Z, r, K_h) = FS - 1$$

such that  $FS$  is obtained through equation (6-17).

In this way, by applying the approach proposed in this document it is possible to compute the bounds of the probability of failure of the rock slope presented in figure 6-9, taking into account the conditions already established for it. The bounds on the probability of failure of this model are computed for different loads  $T \in [0, 700]$  kN, in intervals of 50 kN.

Results of this procedure are presented in figure 6-10 and table 6-5. From these results, it is evident the need for an anchor, or other stabilization measures on the slope since the bounds on the probability of failure are very high when the anchor is not used or when a low load is transmitted by it. As expected, as the load on the anchor increases, both bounds of the probability of failure decrease.

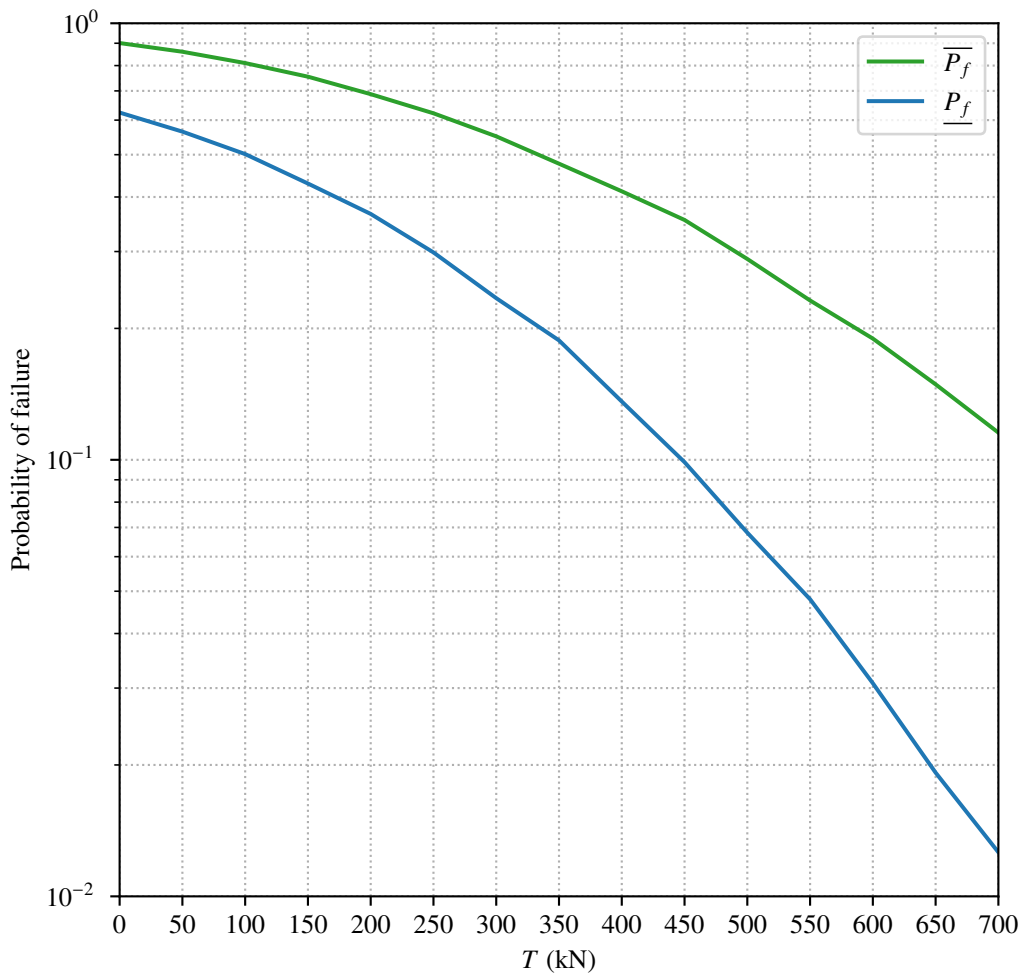
Furthermore, it is worth noting again that as the probability of failure of the slope decreases, the gap between the bounds of the probability of failure increases. This behavior had already been observed in the previous example of this section, and it is mainly due to epistemic uncertainties in the definition of the basic variables of the model. In this way, the impact of epistemic uncertainties becomes more noticeable when high reliability is desired in the geotechnical model, since  $\underline{P}_f$  and  $\overline{P}_f$  can even differ by several orders of magnitude.

Therefore, if models with high reliability are desired, it is necessary to collect more and better information in order to reduce the final uncertainty in the results. This is how the methodology proposed in this thesis demonstrates its usefulness by quantifying this final uncertainty.

With the proposed approach, designers obtain the tools to, at their discretion, define if the uncertainties given by the bounds of the probability of failure are acceptable or if more information is required to reduce the gap between  $\underline{P}_f$  and  $\overline{P}_f$ . For the model studied in this section, the designers could conclude, for example, that a difference of one order of magnitude between  $\underline{P}_f$  and  $\overline{P}_f$  when  $T = 600$  kN is unacceptable, and therefore more information is needed in order to ensure good reliability of the rock slope.

**Table 6-5:** Example 3 of the proposed approach: results of the bounds on the probability of failure in the rock slope from figure 6-9, considering values of  $T$  within [0, 700] kN in intervals of 50 kN.

$T$ (kN)	$\underline{P}_f$	$\overline{P}_f$	$T$ (kN)	$\underline{P}_f$	$\overline{P}_f$
0	$6.244 \times 10^{-1}$	$9.009 \times 10^{-1}$	400	$1.360 \times 10^{-1}$	$4.122 \times 10^{-1}$
50	$5.646 \times 10^{-1}$	$8.606 \times 10^{-1}$	450	$9.878 \times 10^{-2}$	$3.544 \times 10^{-1}$
100	$5.016 \times 10^{-1}$	$8.106 \times 10^{-1}$	500	$6.810 \times 10^{-2}$	$2.884 \times 10^{-1}$
150	$4.294 \times 10^{-1}$	$7.545 \times 10^{-1}$	550	$4.793 \times 10^{-2}$	$2.318 \times 10^{-1}$
200	$3.658 \times 10^{-1}$	$6.881 \times 10^{-1}$	600	$3.078 \times 10^{-2}$	$1.896 \times 10^{-1}$
250	$2.986 \times 10^{-1}$	$6.220 \times 10^{-1}$	650	$1.921 \times 10^{-2}$	$1.489 \times 10^{-1}$
300	$2.346 \times 10^{-1}$	$5.508 \times 10^{-1}$	700	$1.262 \times 10^{-2}$	$1.154 \times 10^{-1}$
350	$1.878 \times 10^{-1}$	$4.769 \times 10^{-1}$			



**Figure 6-10:** Example 3 of the proposed approach: evolution on the bounds of the probability of failure in the rock slope from figure 6-9, considering different load  $T \in [0, 700]$  kN transmitted by an anchor to the sliding plane.

## 7 Conclusions and final considerations

Throughout this document, the issue of uncertainty in geotechnical engineering was addressed in a broad and comprehensive manner. Two major sources of uncertainty were identified: aleatory uncertainty and epistemic uncertainty. Aleatory uncertainty refers to the inherent randomness of phenomena, and it is manifested in geotechnical engineering through the spatial and temporal variability of materials and their properties, as well as, the loads and other external agents that act on them; this uncertainty is irreducible and is characterized through probability theory. On the other hand, epistemic uncertainty refers to the lack of knowledge, which is manifested in geotechnical engineering through limited information or lack of certainty in the models. This uncertainty can be reduced, but it requires obtaining more and better information, which is usually not achieved in geotechnics due to limited budgets or time. These two great sources of uncertainty play a key role in the stability of geotechnical structures and interventions.

However, it is evidenced that the approach with which geotechnical uncertainties have mostly been handled corresponds to a purely deterministic approach. Thus, uncertainties are never characterized or propagated through the models. On the contrary, uncertainties are obviated and, in order to guarantee the stability of the structures, factors of safety are employed. Factors of safety, although are the result of years of experience, may be too excessive, involve many approximations, and in no case provide information on the reliability as a function of uncertainties.

In order to consider uncertainties in geotechnical models, it is necessary to resort to another type of analysis: *the reliability analysis*. Reliability analysis seeks to characterize all the uncertainties, propagate them through the models, and quantify their impact on the safety by estimating a probability of failure  $P_f$ . However, in this master's thesis three fundamental gaps were identified in the reliability analyses commonly carried out in geotechnical engineering:

1. Reliability analyses have traditionally been based on probability theory. As previously stated, with this theory it is possible to characterize and handle aleatory uncertainty; nonetheless, probability theory fails in handling epistemic uncertainty. In this way, the reliability analyses commonly performed in geotechnics only consider aleatory uncertainty. Unfortunately, in geotechnical engineering, where information is so scarce, epistemic uncertainty has a large component and impact on designs and models. Not considering epistemic uncertainty is a great disadvantage and can be risky, since epistemic uncertainty has a role of equal importance (or in some cases of more importance) than aleatory uncertainty.

2. In a probabilistic approach, the probability of failure  $P_f$  is computed by means of equation (2-5). This equation rarely has an analytical solution, so specialized methods must be used to approximate it. Among all the methods that exist for its approximation, simulation methods stand out, and in particular Monte Carlo simulation MCS, which is widely employed for conducting reliability analyses in geotechnics. However, MCS has a great disadvantage: its accuracy carries out to a computational cost that, in many cases, is excessive. Monte Carlo simulation falters when the failure region, associated with the tails of the probability distributions, is too small. In this case, Monte Carlo simulation requires an excessive number of runs to guarantee an acceptable accuracy, since otherwise, its estimates would be misleading.
3. Due to lack of theoretical support or to simplify its calculation, equation (2-5) is subjected to approximations and hypotheses that in many cases are not well-founded. One of the major approximations has to do with the dependence of the basic variables of the model, and therefore with the multivariate probability function that groups all of them. In geotechnics, the practice of assuming independence or gaussian dependence between basic variables is widely extended. However, this assumption is seldom validated or verified, so it can lead to estimates of the probability of failure far from reality.

Fortunately, currently there exist theories and methodologies in concordance with the latest postulates of the state-of-the-art that, although have not been widely implemented for reliability analysis in geotechnical engineering, serve to fill the aforementioned gaps. In particular, this document presents random sets theory, subset simulation algorithm, and copula theory. Random sets theory establishes a framework for the analysis of aleatory and epistemic uncertainties, in the light of which the probabilities of failure of a model are not obtained as a unique value but as lower and upper bounds,  $\underline{P}_f$  and  $\overline{P}_f$ , that converge to a unique value when the epistemic uncertainty disappears; on the other hand, subset simulation algorithm serves to obtain efficient but accurate estimations of  $P_f$ ; this by introducing intermediate failure events in the domain of the basic variables; finally, copula theory provides a framework for the characterization and handling of dependence among basic variables, as well as for the construction of a great variety of multivariate distribution functions in a flexible manner.

The fundamental contribution of this thesis was to apply the aforementioned three theories into a reference framework for carrying out reliability analyses in geotechnical engineering. The proposed approach overcomes the major disadvantages of the purely probabilistic reliability analyses, given that: it is efficient and accurate thanks to the subset simulation algorithm; it allows characterizing dependence and having a wide range of dependence structures to model the basic variables, which is translated into a wide range of multivariate models, thanks to copula theory; it allows to take into account aleatory and epistemic uncertainties in reliability analyses, so that the impact of randomness and lack of information in the estimates of the probability of failure can be evaluated.

Next, some comments and conclusions for each of the methods studied in this document are pre-

sented. Additionally, further conclusions on the results and applicability of the proposed approach for reliability analysis in geotechnics are presented. This thesis ends with some comments about open issues and directions for future research.

## 7.1 Comments on the use of subset simulation in geotechnics

Subset simulation was introduced, in a theoretical but also practical way, in chapter 3 of this document. Also, a keystone in this method, the Markov Chains and Markov Chain Monte Carlo methods, were also presented in chapters 2 and 3, respectively. Finally employability of subset simulation algorithm in geotechnical engineering was evaluated through two practical examples. Based on this postulates, the following conclusions are obtained:

1. By employing subset simulation, the issue of estimating a small probability associated with a rare event is solved by expressed it as the product of conditional probabilities associated with intermediate events, which are more frequent. Informally speaking, subset simulation redirects the simulation to the failure regions instead of spending efforts in the safe regions.
2. For the task of estimating the conditional probabilities associated with the intermediate failure events, subset simulation employs an algorithm based in Markov chain Monte Carlo MCMC methods for generating conditional samples in the intermediate failure regions. In particular, two major algorithms for constructing Markov Chains were presented in this document: the Metropolis algorithm MA and the Metropolis-Hastings algorithm. These two algorithms are the base of the Modified Metropolis algorithm MMA (or the modified Metropolis-Hastings algorithm MMHA), which is employed by SubSim for populating the intermediate failure regions.
3. The modified Metropolis algorithm MMA overcomes some of the disadvantages of other MCMC methods, e.g. Metropolis or Metropolis-Hastings algorithms. Specifically: (1) in the MMA the initial values of the chains are selected implicitly and they will always be adequate for the region to be populated; (2) the type of proposal distribution does not affect the efficiency of the algorithm (although its spreads does); (3) the MMA has the property of perfect sampling (seeds are already distributed according to the target distribution), and hence MMA does not need a burn-in period; (4) MMA does not suffer ergodicity problems, since the algorithm runs several chains at the same time, and then it is expected that the isolated failure regions are populated.
4. The stationarity of the chain generated by the MMA must be guaranteed in order to improve the convergence of the subset simulation algorithm. Furthermore, it is also necessary to avoid auto-correlation on the samples generated through MMA. The spread of the proposal distribution has a great impact on these properties, so it must be defined wisely. Furthermore,

there exists some techniques for evaluating the stationarity and the auto-correlation of the chains generated by the MMA, such as the percentage of acceptance-rejection or the Geweke test, which were presented in this document. Other techniques such as the thinning or lag period can also be implemented.

5. Subset simulation algorithm demonstrated to be robust to the dimension of the models and to the complexity of the limit state functions. Furthermore, it overcomes one of the major drawbacks of the widely employed Monte Carlo simulation: the efficiency. It was demonstrated through theory and practical examples that when probabilities of failure are small, subset simulation algorithm requires much fewer runs of the models, and much less time, than the commonly used Monte Carlo simulation, this without sacrificing accuracy.
6. Subset simulation algorithm proves to be especially suitable to populate and characterize tails of the distributions, where failure regions are commonly located, specially in rare events. This same task would carry a much higher computational cost with the traditional Monte Carlo simulation, particularly when the failure regions are small.
7. Although the advantages in efficiency of subset simulation algorithm were demonstrated in this document by means of a pair of basic examples, the truth is that these advantages are much more noticeable when the complexity of the models increase. In these cases, the difference in time between SubSim and MCS will of several order of magnitude.
8. By taking advantage of the efficiency of the subset simulation algorithm, it is possible to further improve the accuracy of the estimates of  $P_f$ . For this purpose, it is enough to run the algorithm several times and average all the results obtained from each simulation. Some authors have used a total of 50 runs for this purpose; however, this value will depend on the particularities of each model and the covariance found in the results.

## 7.2 Comments on the dependence and copula theory in geotechnics

Copula theory and concepts of dependence in geotechnical engineering were introduced from a theoretical and a practical perspective in chapter 4 of this document. Furthermore, uncertainties in the construction and definition of copula functions, as well as their impact on models, were also included in the state-of-the-art presented in this master's thesis. Finally, the role that copula functions have in geotechnical reliability assessments was also evaluated. Based on this work, the following conclusions can be drawn:

1. Dependence plays a key role in geotechnical engineering and, therefore, it should not be taken lightly, much less overlooked. Dependence on geotechnical parameters will impact, directly or indirectly, the constructed models in a significant way.

2. The practice of geotechnics is full of equations that seek to correlate different parameters and properties of soils and rocks. However, using these equations leads to many uncertainties in the estimates sought and, therefore, in the models built.
3. It is necessary to migrate, in geotechnical engineering, from the concept of correlation to the concept of dependence, which is much broader and, in fact, encompasses correlations. Starting to consider dependence will allow a much better understanding of how geotechnical parameters are related.
4. Copula theory is especially suited for modeling dependence. By means of this theory, a multivariate distribution function can be separated into marginal distributions and a function that groups them together, known as the copula function. This fact is quite useful since it opens up a fairly wide range of possibilities when it comes to modeling multivariate distribution functions and studying the dependence structures between random variables.
5. It is quite common in geotechnics to find that soils and rocks parameters, and in general all random variables, are related in pairs. In many cases, it is enough to use bivariate copulas to model the pairs of random variables. If it is necessary to build models of more variables, they could also be constructed with the conventional copula theory, although this has certain disadvantages related to the assignment of the same dependence structure for all the variables, and that the range of available copula functions is reduced. In this case, it would be much more convenient to make use of vine copulas.
6. There is no ideal copula to model dependence on geotechnics. Actually, there are multiple copula functions in the literature, of which a reduced set must be selected based on the dependence characteristics that we must model. In any case, each copula must be adjusted to the information and the most suitable copula must be chosen through a goodness-of-fit test. However, there will always be some uncertainties in the construction and selection of a copula function, especially when the information is limited.
7. There will always be uncertainties in the construction and selection of copula functions for modeling dependencies in geotechnical engineering. These uncertainties will be greater as the quality and quantity of the information decreases. Thus, the uncertainties will be transmitted to the created models and the results obtained from them. Therefore, it is necessary to build and select copula functions responsibly, in order to obtain models that truthfully represent reality.
8. When conducting geotechnical reliability analysis, different copula functions will lead to different results in the computed probability of failure. As it was demonstrated, probabilities of failure obtained from different copula functions can vary by several orders of magnitude, especially when  $P_f$  is small. This is one more example of the impact of epistemic uncertainty, and it highlights the importance of selecting copulas responsibly. More research should be carried out on the correct construction and selection of copula functions to model

geotechnical variables.

9. Gaussian dependence structure, commonly assumed in geotechnical probability models without any validation, is not always the most suitable copula function. In reality, this dependence structure is only one of many that can be employed in light of copula theory. In the case of reliability analyses, it was shown that Gaussian copula may lead to overestimations of the probability of failure, which becomes too conservative, especially when the probability of failure is small.
10. Currently, Monte Carlo simulation is one of the best alternatives to carry out reliability analyses in geotechnics within the framework of copula theory. It is important to be careful with the number of simulations employed in this technique since an insufficient number would lead to incorrect reliability estimates.
11. Currently, there are simulation techniques, much more advanced than Monte Carlo simulation, which could be used together with copula theory. In particular, subset simulation is a modern technique that allows to calculate small probabilities of failure in a much more efficient way than Monte Carlo simulation. Some studies have already been carried out that integrate subset simulation with copulas. Nonetheless, in this thesis it was demonstrated how these two tools can be incorporated.

### 7.3 Comments on the applicability of random sets theory in geotechnical engineering

Chapter 5 introduced a methodology for the characterization and handling of both aleatory and epistemic uncertainties: *random sets theory*. The applicability of this theory in geotechnical engineering was demonstrated through practical examples. Based on this chapter, the following conclusions and comments can be drawn:

1. Random sets theory manages to integrate aleatory and epistemic uncertainties into a single methodology of analysis. This theory constitutes a generalization of probability theory, where instead of using random variables, random sets are employed, which are basically set-valued random variables. In this regard, when epistemic uncertainty disappears, a random set collapses to a random variable.
2. Through random sets theory it is possible to manipulate different structure of information. In particular, possibility distributions, probability boxes, families of intervals, and CDFs, were introduced in this work, and their relationship with random sets theory was exposed. With these four structures of information it is possible to describe both aleatory and epistemic uncertainties in the definition of parameters and variables in geotechnics. Additionally, it is also possible to include in the modeling the estimates of parameters and values expressed by

means of linguistic terms or intervals, given by reliable sources such as highly experienced geotechnical engineers. This last type of information is quite valuable, and thanks to random sets theory it can be included in the modeling.

3. It is possible to combine different sources information, which would be done to improve the approximations of the structures of information. An example of a methodology of combination of evidence is the Dempster rule of combination, which was introduced in this document. Nonetheless, combination of evidence lead to several approximations, specially when there is strong conflicts between sources. This issue is an open problem in the state-of-the-art that needs more research.
4. In the light of aleatory and epistemic uncertainties, random sets theory states that the probability of an event  $F$ , i.e.  $P_X(F)$ , cannot be computed as an exact value, but rather as an interval with an upper bound and a lower bound. In this regard, a geotechnical model with aleatory and epistemic uncertainties in the definition of its parameters will lead to an upper probability of failure  $\overline{P}_f$  and a lower probability of  $\underline{P}_f$ . As the epistemic uncertainty in the model decreases, the gap between  $\overline{P}_f$  and  $\underline{P}_f$  also decreases, and both bounds converge to  $P_f$  when the epistemic uncertainty approaches to a null value.
5. Using copulas, it is possible to include dependence when studying the association between different random sets.
6. The issue of propagating both aleatory and epistemic uncertainties in the models was addressed through the extension principle. In particular, four different methods were introduced for this purpose, namely the optimization method, the sampling method, the vertex method and the method of function approximation. Each of these methods has its advantages and disadvantages. For instance, the optimization method and the sampling method (with an adequate number of samples) can be very accurate but at the same time computationally expensive; on the other hand, the vertex method and the functions approximation method can be efficient but involve approximation in their estimates. Selection of the most suitable method will depend on the particular conditions of each problem.
7. The discretization of a finite random set has a great impact on the results of the modeling. A fine discretization can guarantee good accuracy but carries a heavy computational cost, while a coarse discretization is efficient but may be imprecise. Therefore, a midpoint between these two extremes must be chosen, which guarantees acceptable accuracy but not excessive computational cost.
8. It was demonstrated that both aleatory and epistemic uncertainties have a great impact in the models. By considering only aleatory uncertainty in the geotechnical models, or in the reliability analyses, one would be incurring in approximations that can lead to misleading results. With random sets theory it is possible to consider both types of uncertainty and evaluate their impact on the models, thus obtaining results according to the state of knowledge. Recall that

imprecise information cannot lead to accurate results, and this is the main motivation for using a theory such as random sets theory.

## 7.4 Conclusions on the proposed methodology for reliability analyses in geotechnics

The major contribution of this master's thesis consists in the integration of subset simulation algorithm, copula theory, and random sets theory, in a comprehensive methodology for the reliability analysis of geotechnical models. This methodology was exposed in chapter 6 of this document, where different examples were also presented for demonstrating its applicability and usefulness. In this sense, the same conclusions obtained for subset simulation, copulas and random sets, are also applicable to the proposed approach. Furthermore, the following additional conclusions are drawn:

1. The proposed approach is a comprehensive methodology for conducting reliability analysis of geotechnical models, in the sense that it considers dependence among the basic variables, both aleatory and epistemic uncertainties, and it is efficient and accurate. This approach overcomes the major drawback of the reliability assessments based uniquely on the probability theory.
2. The proposed approach employs three different spaces for the development of reliability assessments, namely  $\mathcal{X}$ -space  $\Omega$ -space, and  $\mathcal{U}$ -space. The  $\mathcal{X}$ -space is the space of the input variables, where the focal elements of the random sets are contained. The  $\Omega$ -space is the space that contains all the dependence information of the model, and it is represented by a copula function. The  $\mathcal{U}$ -space is the standard gaussian space, where the subset simulation algorithm is conducted. It was explained in this thesis how these three spaces are interrelated and how some transformations can be carried out to move from one of these spaces to the others.
3. It was demonstrated how different structures of information from the  $\mathcal{X}$ -space can be indexed in the  $\Omega$ -space. In particular, this explanation was conducted for possibility distributions, probability boxes, families of intervals, and CDFs. This indexing allows the introduction of an innovative concept: infinite random sets. Instead of discretizing a finite random sets, infinite random sets propose a sampling from the  $\Omega$ -space for then using the indexing in order to relate these samples with focal elements in the  $\mathcal{X}$ -space. This operation allows to overcome drawbacks of the discretizations that are carried out for finite random sets.
4. The  $\Omega$ -space allows a direct and dynamic handling of dependence. Recall that the structure of dependence of this space is defined by the copula function that represents the relationship between the basic variables.

5. With the proposed approach, the upper and lower bounds on the probability of failure of geotechnical models are obtained,  $\overline{P}_f$  and  $\underline{P}_f$ , respectively. The gap between these two bounds is directly related to the epistemic uncertainties in the model. The greater the epistemic uncertainty, the greater the gap expected between these two bounds. This gap evidences the danger of neglecting epistemic uncertainty just for the sake of obtaining a unique value of the probability of failure. Furthermore, it is found that the epistemic uncertainties have a greater impact when the expected probabilities of failure are small; in this case, it is seen that, in the selected examples, the gap between  $\overline{P}_f$  and  $\underline{P}_f$  increases as the probability of failure decreases.
6. The proposed methodology is not intended to invalidate the use of the factor of safety employed for evaluating the stability of geotechnical designs. On the contrary, the proposed approach allows a full understanding of the uncertainties behind each model and can be complemented with the use of factors of safety. With the present methodology, the geotechnical engineers will have the tools to evaluate if the probability of failure in their models is acceptable, or if on the contrary, it should be reduced to have greater reliability.

## 7.5 Open issues and directions for future research

From the work carried out in this document, it is possible to identify some ideas, issues, or problems that require further investigation in the future. These are listed below:

1. Combination of evidence in imprecise probability theory and random sets theory is still an open issue in the state-of-the-art. There are some developments in relation to this issue, such as the Dempster's rule, which was introduced in this document. Nonetheless, up to now, there is no a definitive method to integrate several sources of information.
2. There are multiple uses that copula theory may have in geotechnical engineering, and those studied here are just some of them. In this way, the concepts presented in this document can be applied in many other contexts of geotechnics. For example, copula theory can be evaluated in the study of dependence between parameters of different constitutive models and failure criteria, it could be employed in many of the already existing correlations (e.g., those of the SPT with different geotechnical parameters), and it might even be employed in the study of the relationships between rains, earthquakes, and landslides, among many other uses. The applications of copulas presented in this document are just a hint of the great potential that copulas have in geotechnics. Thus, there is still much to investigate in relation to copula theory and its uses in geotechnical engineering.
3. Vine copula theory has great potential that should be explored in future works. This theory allows to generate high-dimensional models through two-dimensional copulas. By using this theory, it is possible to assign particular dependence structures for each pair of random vari-

ables and, as bivariate copula functions are employed, construct flexible high-dimensional dependence models.

4. Thanks to the subsim algorithm, the proposed approach reduces the number of focal element images needed to compute  $\underline{P}_f$  and  $\overline{P}_f$ . Nonetheless, although this number is considerable reduced, it is still a large number that should be diminished.

# Bibliography

- Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198. (Cited on pages 67 and 68.)
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on Automatic Control*, 19(6):716–723. (Cited on page 96.)
- Alonso, E. E. (1976). Risk analysis of slopes and its application to slopes in canadian sensitive clays. *Geotechnique*, 26(3):453–472. (Cited on page 98.)
- Alvarez, D. A. (2006). On the calculation of the bounds of probability of events using infinite random sets. *International Journal of Approximate Reasoning*, 43(3):241–267. (Cited on pages 154, 172, 176, and 178.)
- Alvarez, D. A. (2007). *Infinite random sets and applications in uncertainty analysis*. PhD thesis, Leopold-Franzens Universitat Innsbruck, Innsbruck, Austria. (Cited on pages 146, 148, 149, 150, 153, 155, 159, 167, 171, 172, 173, 174, 175, 176, and 178.)
- Alvarez, D. A. (2009). A monte carlo-based method for the estimation of lower and upper probabilities of events using infinite random sets of indexable type. *Fuzzy Sets and Systems*, 160(3):384–401. (Cited on pages 172, 176, and 178.)
- Alvarez, D. A. and Hurtado, J. E. (2014). An efficient method for the estimation of structural reliability intervals with random sets, dependence modeling and uncertain inputs. *Computers & Structures*, 142:54–63. (Cited on pages 171, 176, and 177.)
- Alvarez, D. A., Hurtado, J. E., and Ramírez, J. (2017). Tighter bounds on the probability of failure than those provided by random set theory. *Computers & Structures*, 189:101–113. (Cited on pages 155 and 182.)
- Alvarez, D. A., Uribe, F., and Hurtado, J. E. (2018). Estimation of the lower and upper bounds on the probability of failure using subset simulation and random set theory. *Mechanical Systems and Signal Processing*, 100:782–801. (Cited on pages 24, 53, 132, 171, 176, and 179.)
- Ameratunga, J., Sivakugan, N., and Das, B. M. (2016). *Correlations of soil and rock properties in geotechnical engineering*. Springer. (Cited on page 71.)

- Ang, A. H.-S. and Tang, W. H. (1984). Probability concepts in engineering planning and design, vol. 2: Decision, risk, and reliability. *John Wiley & Sons, Inc.* (Cited on pages 8, 11, 29, and 131.)
- Ang, A. H.-S. and Tang, W. H. (2007). *Probability concepts in engineering planning and design: Emphasis on application to civil and environmental engineering*. Wiley. (Cited on pages 8, 71, 75, 76, 77, 79, 80, 91, 93, 94, 102, and 116.)
- Ash, R. B. (2008). *Basic probability theory*. Courier Corporation. (Cited on page 17.)
- Au, S. K. and Beck, J. (2003). Subset simulation and its application to seismic risk based on dynamic analysis. *Journal of Engineering Mechanics*, 129(8):901–917. (Cited on page 47.)
- Au, S.-K. and Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277. (Cited on pages 30, 31, 34, 43, 44, 45, 47, 119, and 132.)
- Au, S. K., Ching, J., and Beck, J. (2007). Application of subset simulation methods to reliability benchmark problems. *Structural Safety*, 29(3):183–193. (Cited on pages 30 and 44.)
- Baecher, G. B. and Christian, J. T. (2005). *Reliability and statistics in geotechnical engineering*. John Wiley & Sons. (Cited on pages 2, 26, and 98.)
- Baker, J. W. and Cornell, C. A. (2006). Correlation of response spectral values for multicomponent ground motions. *Bulletin of the Seismological Society of America*, 96(1):215–227. (Cited on page 115.)
- Baker, J. W. et al. (2007). Correlation of ground motion intensity parameters used for predicting structural and geotechnical response. In *Tenth International Conference on Application of Statistics and Probability in Civil Engineering*, volume 8. Citeseer. (Cited on page 115.)
- Baker, J. W. and Jayaram, N. (2008). Correlation of spectral acceleration values from NGA ground motion models. *Earthquake Spectra*, 24(1):299–317. (Cited on page 115.)
- Barbe, P., Genest, C., Ghoudi, K., and Rémillard, B. (1996). On Kendall's process. *Journal of Multivariate Analysis*, 58(2):197–229. (Cited on page 94.)
- Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1):245–268. (Cited on pages 67 and 68.)
- Bedford, T. and Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068. (Cited on pages 66 and 67.)

- Beer, M., Zhang, Y., Quek, S. T., and Phoon, K. K. (2013). Reliability analysis with scarce information: Comparing alternative approaches in a geotechnical engineering context. *Structural Safety*, 41:1–10. (Cited on page 56.)
- Benjamin, J. R. and Cornell, A. C. (1981). *Probability, Statistics, and Decision for Civil Engineers*. McGraw-Hill. (Cited on page 17.)
- Bernardini, A. and Tonon, F. (2010). *Bounding uncertainty in civil engineering: theoretical background*. Springer Science & Business Media. (Cited on pages 144, 146, 148, and 149.)
- Bertoluzza, C., Gil, M. A., and Ralescu, D. A. (2002). *Statistical modeling, analysis and management of fuzzy data*, volume 87. Physica Verlag. (Cited on page 149.)
- Blockley, D. (1999). Risk based structural safety methods in context. *Structural Safety*, 21(4):335 – 348. (Cited on page 144.)
- Blockley, D. I. (1980). *The nature of structural design and safety*. Ellis Horwood Chichester. (Cited on page 144.)
- Blyth, F. G. H. and De Freitas, M. (2017). *A geology for engineers*. CRC Press. (Cited on page 1.)
- Briaud, J.-L. (2007). Spread footings in sand: load settlement curve approach. *Journal of Geotechnical and Geoenvironmental Engineering*, 133(8):905–920. (Cited on page 107.)
- Burnham, K. P. and Anderson, D. R. (2002). A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed.* Springer, New York. (Cited on page 97.)
- Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304. (Cited on pages 96 and 97.)
- Chao, X. and Lin-de, Y. (1998). Test of goodness of fit of random variables and Bayesian estimation of distribution parameters. *Journal of Tongji University*, 26(3):340–344. (Cited on page 105.)
- Chapra, S. C. et al. (2012). *Applied numerical methods with MATLAB for engineers and scientists*. New York: McGraw-Hill. (Cited on page 25.)
- Chen, L., Singh, V. P., and Guo, S. (2013). Measure of correlation between river flows using the copula-entropy method. *Journal of Hydrologic Engineering*, 18(12):1591–1606. (Cited on page 57.)
- Cheng, Y., Du, J., and Ji, H. (2020). Multivariate joint probability function of earthquake ground motion prediction equations based on vine copula approach. *Mathematical Problems in Engineering*, 2020. (Cited on pages 116 and 117.)

- Cherubini, C. (2000). Reliability evaluation of shallow foundation bearing capacity on  $c - \phi$  soils. *Canadian Geotechnical Journal*, 37(1):264–269. (Cited on page 98.)
- Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons. (Cited on page 57.)
- Chin, F. K. (1970). Estimation of the ultimate load of piles from tests not carried to failure. In *Proceedings, 2nd Southeast Asian Conference on Soil Engineering*, Singapore. (Cited on page 109.)
- Cornell, C. A. (1968). Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58(5):1583–1606. (Cited on page 115.)
- Couso, I., Moral, S., and Walley, P. (1999). Examples of independence for imprecise probabilities. In *Proceedings of 1st International Symposium on Imprecise Probabilities and Their Applications*, volume 99, pages 121–130. (Cited on page 157.)
- Crespo, L. G., Kenny, S. P., and Giesy, D. P. (2013). The NASA Langley multidisciplinary uncertainty quantification challenge. In *16th AIAA Non-Deterministic Approaches Conference*, page 1347. (Cited on page 153.)
- Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*. (Cited on page 66.)
- Davison, M. (1972). High-capacity piles. In *Proceedings, Lecture Series, Innovations in Foundation Construction*, Chicago. ASCE, Illinois Section. (Cited on page 109.)
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38(2):325–339. (Cited on pages 145, 149, 150, 156, and 171.)
- Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112. (Cited on page 56.)
- Der Kiureghian, A. and Liu, P.-L. (1986). Structural reliability under incomplete probability information. *Journal of Engineering Mechanics*, 112(1):85–104. (Cited on pages 26, 72, and 73.)
- Dithinde, M., Phoon, K., De Wet, M., and Retief, J. (2011). Characterization of model uncertainty in the static pile design formula. *Journal of Geotechnical and Geoenvironmental Engineering*, 137(1):70–85. (Cited on pages 106, 107, 108, 109, 110, 113, 123, 124, 125, and 127.)
- Ditlevsen, O. and Madsen, H. O. (1996). *Structural reliability methods*, volume 178. Wiley New York. (Cited on pages 4, 8, 20, 21, 65, and 179.)
- Dong, W. and Shah, H. C. (1987). Vertex method for computing functions of fuzzy variables. *Fuzzy sets and Systems*, 24(1):65–78. (Cited on page 160.)

- Dubois, D. and Prade, H. (1991). Random sets and fuzzy interval analysis. *Fuzzy sets and Systems*, 42(1):87–101. (Cited on pages 145, 151, and 158.)
- Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000). Which copula is the right one? *SSRN Electronic Journal*. (Cited on pages 88, 91, 93, and 101.)
- Dutfoy, A. and Lebrun, R. (2009). Practical approach to dependence modelling using copulas. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 223(4):347–361. (Cited on pages 76, 139, and 140.)
- Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer. (Cited on pages 122 and 123.)
- Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependence with copulas and applications to risk management. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 14. (Cited on pages 58, 61, 76, 77, 80, 81, 82, 83, 85, 100, 103, 110, 134, 178, and 180.)
- Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223. (Cited on pages 58, 59, and 79.)
- Fan, S. (1989). A new extracting formula and a new distinguishing means on the one variable cubic equation. *Nat. Sci. J. Hainan Teach. Coll*, 2(2):91–98. (Cited on page 130.)
- Fellin, W. and Oberguggenberger, M. (2012). Robust assessment of shear parameters from direct shear tests. *International Journal of Reliability and Safety*, 6(1-3):49–64. (Cited on page 98.)
- Fenton, G. A. and Griffiths, D. (2003). Bearing-capacity prediction of spatially random  $c$ - $\phi$  soils. *Canadian Geotechnical Journal*, 40(1):54–65. (Cited on page 98.)
- Ferson, S. (2002). *RAMAS Risk Calc 4.0 software: risk assessment with uncertain numbers*. CRC Press. (Cited on page 152.)
- Ferson, S., Kreinovich, V., Grinzburg, L., Myers, D., and Sentz, K. (2003). Constructing probability boxes and dempster-shafer structures. Technical report, Sandia National Lab. Albuquerque. (Cited on pages 152, 153, 154, 156, and 167.)
- Fetz, T. and Oberguggenberger, M. (2004). Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. *Reliability Engineering & System Safety*, 85(1-3):73–87. (Cited on pages 157 and 158.)
- Forrest, W. S. and Orr, T. L. (2010). Reliability of shallow foundations designed to Eurocode 7. *Georisk*, 4(4):186–207. (Cited on page 55.)

- Fredlund, D. G. and Krahn, J. (1977). Comparison of slope stability methods of analysis. *Canadian Geotechnical Journal*, 14(3):429–439. (Cited on page 181.)
- Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25. (Cited on page 81.)
- Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42. (Cited on page 41.)
- Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368. (Cited on pages 88, 91, 92, and 94.)
- Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283. (Cited on pages 85 and 90.)
- Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213. (Cited on pages 93 and 94.)
- Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043. (Cited on page 94.)
- Ghosh, S. (2010). Modelling bivariate rainfall distribution and generating bivariate correlated rainfall data in neighbouring meteorological subdivisions using copula. *Hydrological Processes*, 24(24):3558–3567. (Cited on page 57.)
- Gilks, W., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis. (Cited on pages 32, 33, 37, 38, and 39.)
- Goda, K. (2010). Statistical modeling of joint probability distribution using copula: application to peak and permanent displacement seismic demands. *Structural Safety*, 32(2):112–123. (Cited on page 57.)
- Goda, K. and Atkinson, G. (2009a). Interperiod dependence of ground-motion prediction equations: A copula perspective. *Bulletin of the Seismological Society of America*, 99(2A):922–927. (Cited on pages 115 and 116.)
- Goda, K. and Atkinson, G. M. (2009b). Probabilistic characterization of spatially correlated response spectra for earthquakes in Japan. *Bulletin of the Seismological Society of America*, 99(5):3003–3020. (Cited on page 116.)
- Goda, K. and Hong, H.-P. (2008). Spatial correlation of peak ground motions and response spectra. *Bulletin of the Seismological Society of America*, 98(1):354–365. (Cited on page 115.)

- Goodman, I. R. and Nguyen, H. T. (2002). Fuzziness and randomness. In *Statistical modeling, analysis and management of fuzzy data*, pages 3–21. Springer. (Cited on page 151.)
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. (Cited on page 35.)
- Hata, Y., Ichii, K., Tsuchida, T., Kano, S., and Yamashita, N. (2008). A practical method for identifying parameters in the seismic design of embankments. *Georisk*, 2(1):28–40. (Cited on page 98.)
- Helton, J. C. (1997). Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Computation and Simulation*, 57(1-4):3–76. (Cited on page 2.)
- Helton, J. C., Johnson, J. D., Oberkampf, W., and Sallaberry, C. J. (2006). Sensitivity analysis in conjunction with evidence theory representations of epistemic uncertainty. *Reliability Engineering & System Safety*, 91(10-11):1414–1434. (Cited on page 146.)
- Hoek, E. (2000). *Practical Rock Engineering*. (Cited on page 162.)
- Hoek, E. and Bray, J. D. (1981). *Rock slope engineering*. CRC Press. (Cited on pages 186, 187, and 191.)
- Huang, D., Yang, C., Zeng, B., and Fu, G. (2014). A Copula-based method for estimating shear strength parameters of rock mass. *Mathematical Problems in Engineering*, 2014. (Cited on pages 101, 102, and 105.)
- Huffman, J. C., Strahler, A. W., and Stuedlein, A. W. (2015). Reliability-based serviceability limit state design for immediate settlement of spread footings on clay. *Soils and Foundations*, 55(4):798–812. (Cited on pages 106, 107, and 108.)
- Huffman, J. C. and Stuedlein, A. W. (2014). Reliability-based serviceability limit state design of spread footings on aggregate pier reinforced clay. *Journal of Geotechnical and Geoenvironmental Engineering*, 140(10):04014055. (Cited on page 108.)
- Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3):587–608. (Cited on page 90.)
- Hurtado, J. E. (2004). *Structural reliability: statistical learning perspectives*, volume 17 of *Lecture Notes in Applied and Computational Mechanics*. Springer Science & Business Media. (Cited on page 20.)
- Hurtado, J. E. and Alvarez, D. A. (2000). Reliability assessment of structural systems using neural networks. In *Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS*, volume 2000. (Cited on page 160.)

- Joe, H. (1996). Families of  $m$ -variate distributions with given margins and  $m(m - 1)/2$  bivariate dependence parameters. *Lecture Notes-Monograph Series*, pages 120–141. (Cited on page 67.)
- Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press. (Cited on pages 80, 84, and 85.)
- Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. Technical report, University of British Columbia. (Cited on pages 91 and 92.)
- Jogdeo, K. (1982). Concepts of dependence. *Encyclopedia of statistical sciences*, 1:324–334. (Cited on pages 14 and 76.)
- Joslyn, C. and Booker, J. M. (2004). Generalized information theory for engineering modeling and simulation. *Engineering Design Reliability Handbook*, 9:1–40. (Cited on page 144.)
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. (Cited on pages 139 and 142.)
- Katafygiotis, L. S. and Zuev, K. M. (2008). Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics*, 23(2-3):208–218. (Cited on page 34.)
- Kazianka, H. and Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic environmental research and risk assessment*, 24(5):661–673. (Cited on page 120.)
- Kazianka, H. and Pilz, J. (2011). Bayesian spatial modeling and interpolation using copulas. *Computers & Geosciences*, 37(3):310–319. (Cited on page 120.)
- Kendall, D. (1974). Foundation of a theory of random sets. *Stochastic geometry*. (Cited on page 145.)
- Klar, A., Aharonov, E., Kalderon-Asael, B., and Katz, O. (2011). Analytical and observational relations between landslide volume and surface area. *Journal of Geophysical Research: Earth Surface*, 116(F2). (Cited on page 130.)
- Klir, G. J. (1995). Principles of uncertainty: What are they? why do we need them? *Fuzzy Sets and Systems*, 74(1):15–31. (Cited on page 146.)
- Kolmogoroff, A. (1941). Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics*, 12(4):461–463. (Cited on page 153.)
- Kottogoda, N. T. and Rosso, R. (2008). *Applied statistics for civil and environmental engineers*. Blackwell Malden, MA. (Cited on pages 8, 11, 75, 94, and 102.)
- Kotz, S. and Drouet, D. (2001). *Correlation and dependence*. World Scientific. (Cited on page 76.)

- Kramer, S. L. et al. (1996). *Geotechnical earthquake engineering*. Pearson Education India. (Cited on page 115.)
- Lambe, T. W. and Whitman, R. V. (1991). *Soil mechanics*, volume 10. John Wiley & Sons. (Cited on pages 1, 19, and 98.)
- Lebrun, R. and Dutfoy, A. (2009a). A generalization of the Nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics*, 24(2):172–178. (Cited on page 74.)
- Lebrun, R. and Dutfoy, A. (2009b). An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, 24(3):312–320. (Cited on page 74.)
- Lee, Y.-F. and Chi, Y.-Y. (2011). Rainfall-induced landslide risk at Lushan, Taiwan. *Engineering Geology*, 123(1-2):113–121. (Cited on page 98.)
- Lemaire, M. (2013). *Structural reliability*. John Wiley & Sons. (Cited on pages 8, 179, and 180.)
- Li, D., Chen, Y., Lu, W., and Zhou, C. (2011). Stochastic response surface method for reliability analysis of rock slopes involving correlated non-normal variables. *Computers and Geotechnics*, 38(1):58–68. (Cited on page 98.)
- Li, D., Tang, X., Zhou, C., and Phoon, K. K. (2012). Uncertainty analysis of correlated non-normal geotechnical parameters using Gaussian copula. *Science China Technological Sciences*, 55(11):3081–3089. (Cited on pages 107, 110, 111, 113, and 115.)
- Li, D. Q., Tang, X. S., Phoon, K. K., Chen, Y. F., and Zhou, C. B. (2013). Bivariate simulation using copula and its application to probabilistic pile settlement analysis. *International Journal for Numerical and Analytical Methods in Geomechanics*, 37(6):597–617. (Cited on pages 88, 107, 110, 111, 112, 113, 115, 123, and 127.)
- Li, D.-Q., Tang, X.-S., Zhou, C.-B., and Phoon, K.-K. (2015a). Characterization of uncertainty in probabilistic model using bootstrap method and its application to reliability of piles. *Applied Mathematical Modelling*, 39(17):5310–5326. (Cited on pages 115, 123, 125, and 127.)
- Li, D. Q., Zhang, L., Tang, X. S., Zhou, W., Li, J. H., Zhou, C. B., and Phoon, K. K. (2015b). Bivariate distribution of shear strength parameters using copulas and its impact on geotechnical system reliability. *Computers and Geotechnics*, 68:184–195. (Cited on pages 99, 128, 133, and 137.)
- Li, H., Lü, Z., and Yuan, X. (2008). Nataf transformation based point estimate method. *Chinese Science Bulletin*, 53(17):2586. (Cited on pages 73 and 74.)
- Li, K. and Lumb, P. (1987). Probabilistic design of slopes. *Canadian Geotechnical Journal*, 24(4):520–535. (Cited on page 98.)

- Liu, P.-L. and Der Kiureghian, A. (1986). Multivariate distribution models with prescribed marginals and covariances. *Probabilistic Engineering Mechanics*, 1(2):105–112. (Cited on page 73.)
- Lizarraga, H. S. and Lai, C. G. (2014). Effects of spatial variability of soil properties on the seismic response of an embankment dam. *Soil Dynamics and Earthquake Engineering*, 64:113–128. (Cited on page 98.)
- Low, B. (2007). Reliability analysis of rock slopes involving correlated nonnormals. *International Journal of Rock Mechanics and Mining Sciences*, 44(6):922–935. (Cited on page 98.)
- Lumb, P. (1970). Safety factors and the probability distribution of soil strength. *Canadian Geotechnical Journal*, 7(3):225–242. (Cited on page 98.)
- Luo, Z., Atamturktur, S., and Juang, C. H. (2013). Bootstrapping for characterizing the effect of uncertainty in sample statistics for braced excavations. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(1):13–23. (Cited on pages 123 and 124.)
- Malevergne, Y., Sornette, D., et al. (2003). Testing the Gaussian copula hypothesis for financial assets dependences. *Quantitative Finance*, 3(4):231–250. (Cited on page 94.)
- Marchant, B. P., Saby, N. P., Jolivet, C. C., Arrouays, D., and Lark, R. M. (2011). Spatial prediction of soil properties with copulas. *Geoderma*, 162(3-4):327–334. (Cited on page 119.)
- Marek, P., Anagnos, T., and Gustar, M. (1996). *Simulation-based reliability assessment for structural engineers*. CRC Press. (Cited on pages 20 and 29.)
- Matheron, G. (1974). *Random Sets and Integral Geometry*. Probability and Statistics Series. Wiley. (Cited on page 145.)
- Matsuo, M. and Kuroda, K. (1974). Probabilistic approach to design of embankments. *Soils and Foundations*, 14(2):1–17. (Cited on page 98.)
- Mayne, P. W. and Poulos, H. G. (1999). Approximate displacement influence factors for elastic shallow foundations. *Journal of Geotechnical and Geoenvironmental Engineering*, 125(6):453–460. (Cited on page 106.)
- McGuire, R. K. (2004). *Seismic hazard and risk analysis*. Earthquake Engineering Research Institute. (Cited on page 115.)
- McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton University Press. (Cited on pages 57, 58, 64, 78, 82, 83, 88, 90, and 103.)
- Melchers, R. E. and Beck, A. T. (2018). *Structural reliability analysis and prediction*. John Wiley & Sons. (Cited on pages 4, 8, 21, 22, 23, 24, 26, 28, 29, and 132.)

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092. (Cited on page 33.)
- Mitchell, J. K., Soga, K., et al. (2005). *Fundamentals of Soil behavior*, volume 3. John Wiley & Sons New York. (Cited on pages 1 and 98.)
- Montgomery, D. C. and Runger, G. C. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons. (Cited on pages 11 and 75.)
- Most, T. and Knabe, T. (2010). Reliability analysis of the bearing failure problem considering uncertain stochastic parameters. *Computers and Geotechnics*, 37(3):299–310. (Cited on pages 123 and 124.)
- Motamedi, M. and Liang, R. Y. (2014). Probabilistic landslide hazard assessment using copula modeling technique. *Landslides*, 11(4):565–573. (Cited on pages 120 and 121.)
- Nasekhian, A. and Schweiger, H. F. (2011). Random set finite element method application to tunnelling. *International Journal of Reliability and Safety*, 5(3-4):299–319. (Cited on page 161.)
- Nataf, A. (1962). Détermination des distributions de probabilités dont les marges sont données. *C. R. Acad Sci*, 225:42–43. (Cited on pages 56 and 72.)
- Naylor, T., Naylor, T., Balintfy, J., Burdick, D., and Chu, K. (1966). *Computer Simulation Techniques*. Wiley. (Cited on page 27.)
- Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media. (Cited on pages 14, 58, 60, 62, 64, 65, 76, 78, 80, 83, 85, 87, 88, 89, 90, 99, 100, 178, and 180.)
- Nguyen, V. and Chowdhury, R. (1984). Probabilistic study of spoil pile stability in strip coal mines—two techniques compared. In *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, volume 21, pages 303–312. Elsevier. (Cited on page 98.)
- Oberguggenberger, M. and Fellin, W. (2002). From probability to fuzzy sets: the struggle for meaning in geotechnical risk assessment. In *Conference Report*, volume 1. Citeseer. (Cited on page 144.)
- Oberguggenberger, M. and Fellin, W. (2004). *The fuzziness and sensitivity of failure probabilities*, pages 33–49. (Cited on page 144.)
- Oberguggenberger, M. and Fellin, W. (2005). Assessing the sensitivity of failure probabilities: a random set approach. In *Safety and Reliability of Engineering Systems and Structures: Proceedings of the 9th International Conference on Structural Safety and Reliability*, pages 1755–1760. (Cited on page 153.)

- Oberguggenberger, M. and Fellin, W. (2008). Reliability bounds through random sets: non-parametric methods and geotechnical applications. *Computers & Structures*, 86(10):1093–1101. (Cited on page 153.)
- Oberkampf, W. L., Tucker, W. T., Zhang, J., Ginzburg, L., Berleant, D. J., Ferson, S., Hajagos, J., and Nelsen, R. B. (2004). Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis. Technical report, Sandia National Laboratories. (Cited on pages 157 and 158.)
- Orr, T. L. (2000). Selection of characteristic values and partial factors in geotechnical designs to Eurocode 7. *Computers and Geotechnics*, 26(3-4):263–279. (Cited on page 55.)
- Parker, C., Simon, A., and Thorne, C. R. (2008). The effects of variability in bank material properties on riverbank stability: Goodwin Creek, Mississippi. *Geomorphology*, 101(4):533–543. (Cited on page 98.)
- Peck, R. B. (1969). Advantages and limitations of the observational method in applied soil mechanics. *Geotechnique*, 19(2):171–187. (Cited on pages 3 and 4.)
- Peschl, G. and Schweiger, H. (2004). Application of the random set finite element method (RS-FEM) in geotechnics. In *Plaxis Bulletin*, volume 19. (Cited on page 161.)
- Peschl, G. M. (2004). *Reliability Analyses in Geotechnics with the Random Set Finite Element Method*. Phd thesis, Technische Universitat Graz, Graz, Austria. (Cited on page 161.)
- Phoon, K., Chen, J., and Kulhawy, F. (2006). Characterization of model uncertainties for augured cast-in-place (ACIP) piles under axial compression. In *Foundation Analysis and Design: Innovative Methods*, pages 82–89. (Cited on pages 107, 108, and 110.)
- Phoon, K., Chen, J.-R., and Kulhawy, F. (2007). Probabilistic hyperbolic models for foundation uplift movements. In *Probabilistic Applications in Geotechnical Engineering*, pages 1–12. (Cited on page 110.)
- Phoon, K.-K. (2008). *Reliability-based design in geotechnical engineering: computations and applications*. CRC Press. (Cited on pages 49 and 110.)
- Phoon, K.-K. (2020). The story of statistics in geotechnical engineering. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 14(1):3–25. (Cited on page 3.)
- Phoon, K.-K. and Ching, J. (2014). *Risk and reliability in geotechnical engineering*. CRC Press. (Cited on pages 4, 85, 99, 101, 102, 103, 128, 130, 131, 132, 134, and 137.)
- Phoon, K.-K., Kulhawy, F. H., and Grigoriu, M. D. (2003). Multiple resistance factor design for shallow transmission line structure foundations. *Journal of Geotechnical and Geoenvironmental Engineering*, 129(9):807–818. (Cited on page 107.)

- Puzrin, A. M., Alonso, E. E., and Pinyol, N. M. (2010). *Geomechanics of failures*. Springer Science & Business Media. (Cited on page 19.)
- Rackwitz, R. (2000). Reviewing probabilistic soils modeling. *Computers and Geotechnics*, 26:199–223. (Cited on page 144.)
- Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media. (Cited on page 46.)
- Robertson, P. and Cabal, K. (2015). *Guide to Cone Penetration Testing For Geotechnical Engineering*. Gregg Drilling & Testing, Inc. (Cited on page 107.)
- Ross, S. (2012). *Simulation*. Knovel Library. Elsevier Science. (Cited on page 28.)
- Rubinstein, R. Y. and Kroese, D. P. (2016). *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons. (Cited on page 28.)
- Schuëller, G., Pradlwarter, H., and Koutsourelakis, P.-S. (2004). A critical appraisal of reliability estimation procedures for high dimensions. *Probabilistic Engineering Mechanics*, 19(4):463–474. (Cited on page 132.)
- Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. (Cited on page 96.)
- Schweiger, H. and Peschl, G. (2005). Reliability analysis in geotechnics with the random set finite element method. *Computers and Geotechnics*, 32:422–435. (Cited on page 161.)
- Schweiger, H. and Peschl, G. M. (2004). Numerical analysis of deep excavations utilizing random set theory. In *Geotechnical Innovations*, pages 277–294. (Cited on page 161.)
- Schweizer, B., Wolff, E. F., et al. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885. (Cited on page 89.)
- Sentz, K., Ferson, S., et al. (2002). *Combination of evidence in Dempster-Shafer theory*, volume 4015. Sandia National Laboratories Albuquerque. (Cited on pages 145, 146, and 156.)
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press. (Cited on pages 145 and 156.)
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231. (Cited on pages 56 and 61.)
- Sklar, A. (1996). Random variables, distribution functions, and copulas: A personal look backward and forward. *Lecture Notes-Monograph Series*, 28:1–14. (Cited on page 61.)

- Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14. (Cited on page 153.)
- Soubra, A.-H. and Mao, N. (2012). Probabilistic analysis of obliquely loaded strip foundations. *Soils and Foundations*, 52(3):524–538. (Cited on page 98.)
- Sriboonchitta, S. and Kreinovich, V. (2018). Why are FGM copulas successful? a simple explanation. *Advances in Fuzzy Systems*, 2018. (Cited on page 99.)
- Staff, M.-W. (2004). *Merriam-Webster's collegiate dictionary*, volume 2. Merriam-Webster. (Cited on page 1.)
- Stansbury, Dustin (2012). MCMC the Metropolis-Hastings sampler. <https://theclevermachine.wordpress.com/2012/10/20/mcmc-the-metropolis-hastings-sampler/>. [Online; accessed 28-December-2020]. (Cited on page 39.)
- Tang, X. S., Li, D. Q., Cao, Z. J., and Phoon, K. K. (2017). Impact of sample size on geotechnical probabilistic model identification. *Computers and Geotechnics*, 87:229–240. (Cited on pages 99, 101, 102, 122, and 123.)
- Tang, X.-S., Li, D.-Q., Chen, Y.-F., Zhou, C.-B., and Zhang, L.-M. (2012). Improved knowledge-based clustered partitioning approach and its application to slope reliability analysis. *Computers and Geotechnics*, 45:34–43. (Cited on page 98.)
- Tang, X. S., Li, D. Q., Rong, G., Phoon, K. K., and Zhou, C. B. (2013). Impact of copula selection on geotechnical reliability under incomplete probability information. *Computers and Geotechnics*, 49:264–278. (Cited on pages 101, 105, 128, 129, 130, 131, 132, 133, 134, 135, 137, 138, and 142.)
- Tang, X. S., Li, D. Q., Zhou, C. B., and Phoon, K. K. (2015). Copula-based approaches for evaluating slope reliability under incomplete probability information. *Structural Safety*, 52(PA):90–99. (Cited on pages 99, 129, 130, 132, 134, 137, 138, 139, 140, 142, and 143.)
- Tarbut, E. J., Lutgens, F. K., Tasa, D., and Linneman, S. (2005). *Earth: an introduction to physical geology*. Pearson/Prentice Hall Upper Saddle River. (Cited on page 1.)
- Terzaghi, K. (1943). Theoretical soil mechanics. *John Wiley and Sons*. (Cited on pages 52 and 53.)
- Terzaghi, K., Peck, R. B., and Mesri, G. (1996). *Soil mechanics in engineering practice*. John Wiley & Sons. (Cited on page 3.)
- Tobutt, D. and Richards, E. (1979). The reliability of earth slopes. *International Journal for Numerical and Analytical Methods in Geomechanics*, 3(4):323–354. (Cited on page 98.)

- Tonon, F. (2004). Using random set theory to propagate epistemic uncertainty through a mechanical system. *Reliability Engineering & System Safety*, 85(1-3):169–181. (Cited on pages 160, 166, 167, and 169.)
- Tonon, F., Bernardini, A., and Mammino, A. (2000a). Determination of parameters range in rock engineering by means of random set theory. *Reliability Engineering System Safety*, 70:241–261. (Cited on pages 150 and 161.)
- Tonon, F., Bernardini, A., and Mammino, A. (2000b). Reliability analysis of rock mass response by means of random set theory. *Reliability Engineering & System Safety*, 70:263–282. (Cited on pages 161 and 164.)
- Tonon, F., Mammino, A., Bernardini, A., et al. (1996). A random set approach to the uncertainties in rock engineering and tunnel lining design. In *ISRM International Symposium-EUROCK 96*. International Society for Rock Mechanics and Rock Engineering. (Cited on page 161.)
- Uribe, F. (2011). Implementation of simulation methods in structural reliability. Master's thesis, Universidad Nacional de Colombia. (Cited on pages 16, 17, 31, 34, and 38.)
- Uzielli, M. and Mayne, P. W. (2011). Serviceability limit state CPT-based design for vertically loaded shallow footings on sand. *Geomechanics and Geoengineering*, 6(2):91–107. (Cited on pages 106, 107, and 108.)
- Uzielli, M. and Mayne, P. W. (2012). Load-displacement uncertainty of vertically loaded shallow footings on sands and effects on probabilistic settlement estimation. *Georisk*, 6(1):50–69. (Cited on pages 106, 107, and 108.)
- Vanmarcke, E. (2010). *Random fields: analysis and synthesis*. World scientific. (Cited on page 117.)
- Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228. (Cited on pages 96 and 97.)
- Walsh, B. (2004). Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, 581. (Cited on pages 17, 33, 37, 38, and 39.)
- Wang, F. and Li, H. (2018). The role of copulas in random fields: Characterization and application. *Structural Safety*, 75:75–88. (Cited on page 119.)
- Wang, J. P., Tang, X. S., Wu, Y. M., and Li, D. Q. (2018). Copula-based earthquake early warning decision-making strategy. *Soil Dynamics and Earthquake Engineering*, 115:324–330. (Cited on page 121.)

- Wang, M.-X., Tang, X.-S., Li, D.-Q., and Qi, X.-H. (2020). Subset simulation for efficient slope reliability analysis involving copula-based cross-correlated random fields. *Computers and Geotechnics*, 118:103326. (Cited on pages 118, 119, and 132.)
- Wang, Z. and Klir, G. (1992). *Fuzzy Measure Theory*. The Language of science. Springer US. (Cited on page 148.)
- Whittle, A. and Davies, R. (2006). Nicoll highway collapse: evaluation of geotechnical factors affecting design of excavation support system. In *International Conference on Deep Excavations*, volume 28, page 30. (Cited on page 19.)
- Wikipedia contributors (2021a). Autocorrelation — Wikipedia, the free encyclopedia. <https://en.wikipedia.org/wiki/Autocorrelation>. [Online; accessed 3-January-2021]. (Cited on page 14.)
- Wikipedia contributors (2021b). Examples of Markov chains — Wikipedia, the free encyclopedia. [https://en.wikipedia.org/wiki/Examples\\_of\\_Markov\\_chains](https://en.wikipedia.org/wiki/Examples_of_Markov_chains). [Online; accessed 5-January-2021]. (Cited on page 17.)
- Wikipedia contributors (2021c). Markov chain — Wikipedia, the free encyclopedia. [https://en.wikipedia.org/wiki/Markov\\_chain](https://en.wikipedia.org/wiki/Markov_chain). [Online; accessed 4-Januray-2021]. (Cited on page 15.)
- Wolff, T. F. (1985). *Analysis and design of embankment dam slopes: a probabilistic approach*. University Microfilms. (Cited on page 98.)
- Wu, X. Z. (2013a). Probabilistic slope stability analysis by a copula-based sampling method. *Computational Geosciences*, 17(5):739–755. (Cited on pages 99, 100, 101, 102, 128, 129, and 132.)
- Wu, X. Z. (2013b). Trivariate analysis of soil ranking-correlated characteristics and its application to probabilistic stability assessments in geotechnical engineering problems. *Soils and Foundations*, 53(4):540–556. (Cited on pages 99, 102, 128, 129, 133, and 138.)
- Wu, X. Z. (2015). Modelling dependence structures of soil shear strength data with bivariate copulas and applications to geotechnical reliability analysis. *Soils and Foundations*, 55(5):1243–1258. (Cited on pages 100, 101, 105, 128, and 138.)
- Wyllie, D. C. (2017). *Rock slope engineering: civil applications*. CRC Press. (Cited on pages 186, 187, and 191.)
- Xu, X., Li, J., Gong, J., Deng, H., and Wan, L. (2016a). Copula-Based Slope Reliability Analysis Using the Failure Domain Defined by the g-Line. *Mathematical Problems in Engineering*, 2016. (Cited on pages 99, 100, 101, 102, 128, 129, 130, 131, 132, 137, 140, and 143.)

- Xu, Y., Tang, X. S., Wang, J. P., and Kuo-Chen, H. (2016b). Copula-based joint probability function for PGA and CAV: a case study from Taiwan. *Earthquake Engineering and Structural Dynamics*, 45(13):2123–2136. (Cited on pages 115 and 116.)
- Xu, Z.-X. and Zhou, X.-P. (2018). Three-dimensional reliability analysis of seismic slopes using the copula-based sampling method. *Engineering Geology*, 242:81–91. (Cited on pages 129, 132, and 137.)
- Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41(2):93–137. (Cited on page 156.)
- Yu, Q. (2006). Slope reliability of embankment dam and its application to engineering practice. Master's thesis, Hohai University, Nanjing, China. (Cited on page 99.)
- Zhang, J., Huang, H. W., Juang, C. H., and Su, W. W. (2014). Geotechnical reliability analysis with limited data: Consideration of model selection uncertainty. *Engineering Geology*, 181:27–37. (Cited on pages 101, 105, 128, 137, 138, 139, 140, and 142.)
- Zhang, L. and Singh, V. (2006). Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering*, 11(2):150–164. (Cited on page 57.)
- Zhang, L. and Singh, V. P. (2019). *Copulas and their applications in water resources engineering*. Cambridge University Press. (Cited on page 87.)
- Zhang, L., Tang, X., and Li, D. (2013). Bivariate distribution model of soil shear strength parameter using copula. *Journal of Civil Engineering and Management*, 30(2):11–17. (Cited on pages 99 and 105.)
- Zhu, H., Zhang, L., Xiao, T., and Li, X. (2017). Generation of multivariate cross-correlated geotechnical random fields. *Computers and Geotechnics*, 86:95–107. (Cited on pages 117, 118, and 119.)
- Zou, Z.-H., Yi, Y., and Sun, J.-N. (2006). Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. *Journal of Environmental Sciences*, 18(5):1020–1023. (Cited on page 143.)
- Zuev, K. M., Beck, J. L., Au, S.-K., and Katafygiotis, L. S. (2012). Bayesian post-processor and other enhancements of subset simulation for estimating failure probabilities in high dimensions. *Computers & structures*, 92:283–296. (Cited on pages 29, 32, 41, 44, 45, 46, 48, and 53.)