Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

MODELING DEPENDENCE IN SIMULATION INPUT: THE CASE FOR COPULAS

Raghu Pasupathy

Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Kalyani Nagaraj

Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

ABSTRACT

We discuss copulas for incorporating dependence in the input distributions to a simulation model. We start by motivating the need for incorporating dependence in the primitive inputs to a simulation. Copulas are then introduced as a convenient and flexible model to incorporate dependence. We rigorously define copulas, introduce some of their basic properties, illustrate popular copula families, and discuss methods for copula estimation. Since this is an introductory tutorial, we have attempted to keep all exposition at a basic mathematical level without omitting important technical details. The oral presentation of this tutorial will include additional discussion on random variate generation, copula inference, and tail dependence.

1 INTRODUCTION

What is a simulation model? A precise answer to this question is difficult and we will not undertake an explanation; instead, for the purposes of this tutorial, we will understand simulation models loosely, as computerized abstractions of the functioning of some physical system of interest. The physical system may be a portfolio of stock prices evolving in time, a large manufacturing plant engaged in the daily production of widgets, or a small city reeling under the spread of an epidemic. Simulating these systems is usually an attempt at faithfully but parsimoniously representing (on a digital computer) the system's functioning with the objective of estimating a well-defined performance measure. For example, in simulating the portfolio of stocks, we may seek the expected revenue at the end of a specified time horizon; or we may simulate the operations of a large production plant in an attempt at estimating the weekly expected throughput.

The examples just listed — evolution of a stock portfolio, the operations of a production plant, or the spreading of an epidemic in a city — are all very different physical contexts but their simulations (or computerized abstraction) are much less so. Nelson (1987) notes in his seminal work that virtually all simulations, explicitly or implicitly, are subsumed by the simple simulation worldview depicted in Figure 1. The worldview consists of five essential components: a set of random seeds X_0, a set of random numbers U obtained from X_0 using the generator $\mathscr{G}_r$, a set of input data V obtained from U through an input model $\mathscr{I}$, a set of output data Y obtained from V using the simulation logic $\mathscr{L}$, and finally the estimator set $\hat{\theta}$ that estimates the true set of performance measures of interest θ and obtained from Y using the statistical estimation procedure S. (For brevity, we will not discuss the other elements of the simulation world-view — see Leemis (2003) for a detailed explanation.)

This tutorial is about the choice of the crucial element $\mathscr{I}$ representing the input model used within the simulation worldview (Figure 1). The input model is usually a probability model that results from applying a process $\mathscr{P}$ to a set of data $\mathscr{D}$ collected from the physical system. So, for the context of simulating the portfolio of stocks evolving in time, $\mathscr{I}$ might represent the joint distributions that simultaneously drive the evolution of stock prices within the portfolio. In this tutorial, we address the question of how to incorporate

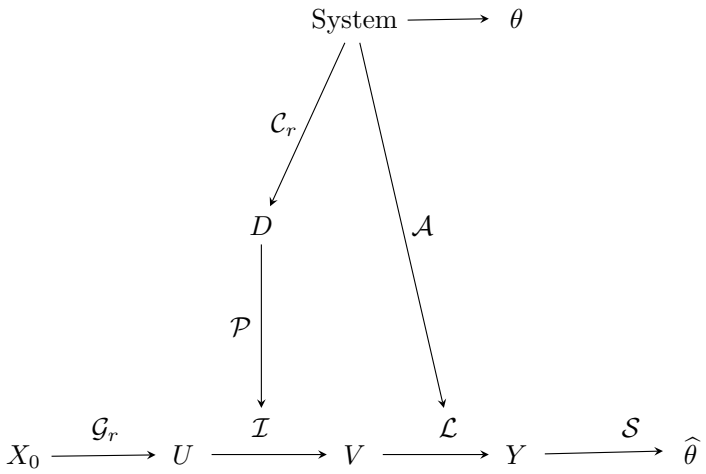


Figure 1: A simulation worldview depicting the random seeds X_0, the random numbers U, the input data V, the output data Y and the estimator(s) $\hat{\theta}$. A "correct" input model $\mathscr{I}$ is crucial to building simulations that faithfully represent the system of interest. This tutorial focuses on incorporating dependence within the input model $\mathscr{I}$.

dependence within the collection of distributions comprising $\mathscr{I}$. Furthermore, the focus is much less on the (marginal) distributions that drive individual stock prices than on their interaction or *dependence*.

Is modeling dependence in simulation input important? Can dependence be ignored so that the distributions that collectively comprise $\mathscr{I}$ in the simulation worldview be modeled independently? The answer to this question is best illustrated using a simple example.

Example. Consider a multiple item production and inventory system that services a multi-product assemble-to-order plant. The company makes and stocks only the components and not the final products, a common setup in the electronics and automobile manufacturing industries (Glasserman and Wang 1998, Xu 1999, Song and Zipkin 2003). Alternatively, the operations of a blood center that collects, stores, and delivers blood and blood products to regional hospitals and clinics may be viewed in a similar vein. Such production and inventory systems are often managed separately from the assembly lines, especially when component manufacturing or procurement times are significantly longer than assembly leadtimes, or in the case of perishable goods, the products' shelf lives. While orders for different products may arrive independently, ignoring the implied dependence between demands on the components comprising a product order can lead to inefficient stocking policies in the inventory system.

For expository purposes, let us assume a two-item production system in which the true demands come from a bivariate NORTA distribution (Law 2007) with exponential marginal distribution functions $D_1 \sim \text{Exp}(1/10), D_2 \sim \text{Exp}(1/90)$, and having correlation $\rho = 0.9$. Figure 2 shows two scatter plots of simulated demand random vectors; the left panel shows generated demands with the correct marginal distribution functions but with dependence ignored, and the right panel shows generated demands from the true distribution where dependence is incorporated. As expected, the scatterplots show distinctly different trends, with demands on the right panel being high or low together, unlike the left panel where there is no discernible relationship.

How does ignoring dependence in the probability model for demand (D_1, D_2) manifest itself in decision-making? To answer this question, suppose we use the demand probability model to make a decision on how much inventory to hold. Specifically, suppose we wish to identify the minimum total inventory level that will ensure that the stockout probability remains below a threshold. Formally, if s_1 and s_2 are the inventory levels with corresponding holding costs c_1 and c_2, and α represents the threshold stockout probability, the optimization problem to identify the optimal inventory levels is given as:

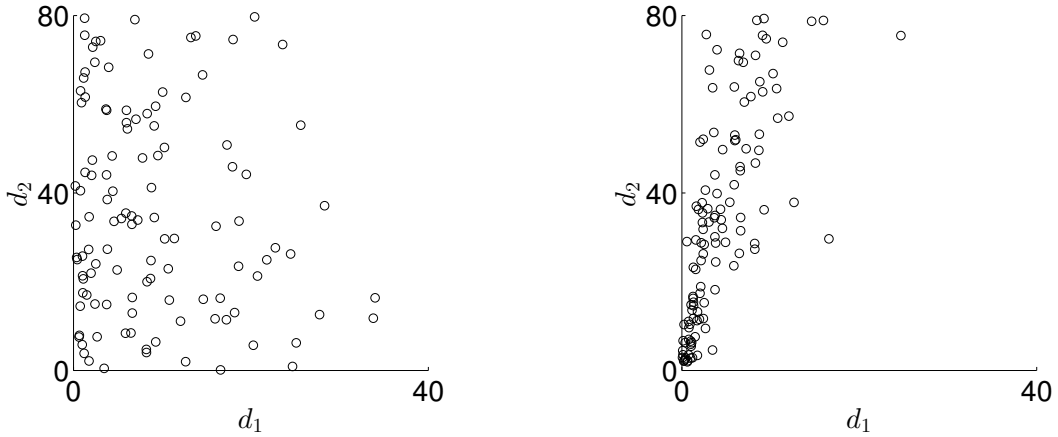


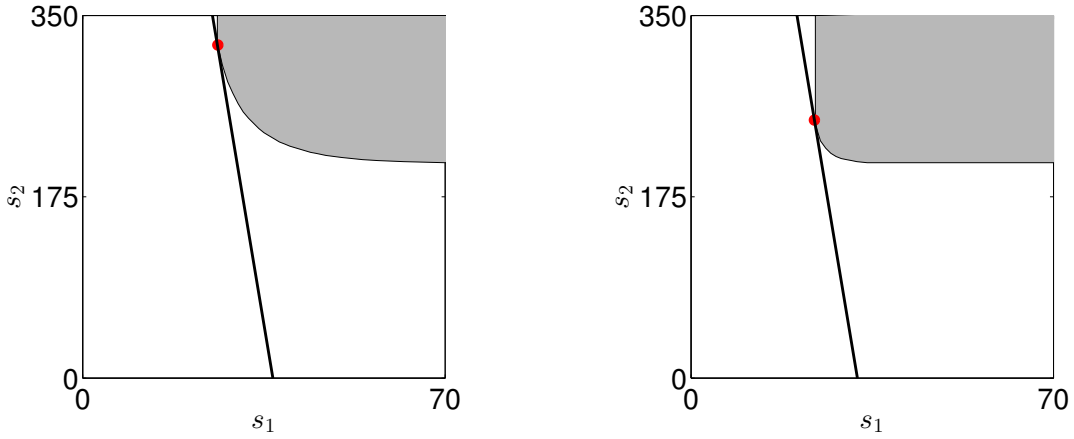
Figure 2: Scatterplots of simulated demands D_1 and D_2 under two probability models: (a) ignore dependence but incorporate the correct marginal distribution functions; and (b) incorporate the correct marginal distribution functions and the dependence structure.

$$\begin{aligned} \text{minimize} \quad & c_1 s_1 + c_2 s_2 \\ \text{subject to} \quad & \Pr\left\{ D_1 > s_1 \bigcup D_2 > s_2 \right\} \leq \alpha. \end{aligned} \tag{1}$$

It is evident from the problem formulation in (1) that the assumed probability model will directly affect the feasible region, potentially introducing errors in the optimal solution. Figure 3 depicts the feasible region and the optimal solution of the problem in (1) for the incorrect and correct probability models. It can be seen that ignoring the dependence in demands D_1 and D_2 results in a nontrivial change in the structure of the feasible region, and consequently, substantially different inventory policies. The numbers shown on the right and left panels of Figure 3 quantify this difference. For instance, the suggested inventory level when ignoring dependence results in an error of almost 30 percent for s_2, with a corresponding 10 percent increase in holding costs. In this example, ignoring dependence seems to have little effect on the choice of s_1. ■

The above example was constructed on a small scale for expository purposes, but delivers an important message: ignoring dependence in simulation input, while convenient, can have a nontrivial impact on the decisions made using the simulation output. In more complicated settings, there is reason to believe that the effect of ignoring dependence can be more severe. Sound theoretical arguments based on the cascading effects of modeling error, along with a number of high-profile negative examples bolster this view.

The idea that modeling dependence in simulation input can be important is clear to most academicians and well-recognized among implementers. Yet, it appears that most existing discrete-event simulation packages at the time of this writing have at best only rudimentary mechanisms to handle such dependence. For example, the package Simio does not yet support methods for generation from general multivariate distributions, although certain specific stochastic processes are supported. Similarly, the popular package ARENA (Rockwell Automation) does not have any inbuilt mechanism for dependent random variate generation; neither does the recently introduced SAS Simulation Studio, although support in the form of interfacing ability with SAS JMP, which has copula modeling capabilities, is provided.



(a) $s^*_{\text{indep}} = (25.6, 333.7), v^*_{\text{indep}} = 1101.5$ (b) $s^*_{\text{dep}} = (24.1, 264.3), v^*_{\text{dep}} = 988.6$

Figure 3: The figure shows the feasible region and the optimal solution for the optimization problem in (1) when (a) the marginal distribution functions are correctly chosen but the dependence ignored; and (ii) both the marginal distribution functions and the correct dependence structure are incorporated.

1.1 Organization of the Tutorial

This tutorial undertakes a basic exposition of copulas as a flexible mechanism for incorporating dependence in simulation input. After motivating copulas in the bivariate context in Section 2, we formally define copulas and some of their key properties in Section 3. This is followed by Section 4 where we describe and illustrate popular copula families. Sections 5 addresses estimation for copulas.

Since this is an introductory tutorial, we have attempted to keep all exposition at a basic mathematical level without omitting important technical details. We do not treat more advanced theoretical and methodological underpinnings of copulas, especially in high dimensions and for heavy tail dependence. We also do not treat process copulas but instead limit ourselves to modeling random vectors. Section 6 provides some concluding remarks.

1.2 Notation and Convention

We will adopt the following notation through the paper. (i) If $\mathbf{x} \in \mathbb{R}^d$ is a vector, then its components are denoted through $\mathbf{x} := (x_1, x_2, \ldots, x_d)$. (ii) If a random variable U is uniformly distributed in the interval $[0,1]$, we write $U \sim U[0,1]$; (iii) We write $F^{-1}(t) = \inf\{s : F(s) \geq t\}$ to represent the generalized inverse of the cumulative distribution function (cdf) F. (iv) If ρ is a $d \times d$ matrix, then $|\rho|$ denotes its determinant; (v) $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard Gaussian distribution function (df) and the standard Gaussian density function respectively; (vi) The L^2-norm of a vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$.

2 INCORPORATING DEPENDENCE IN SIMULATION INPUT

For ease of exposition, let us consider the bivariate context. Suppose we wish to model a bivariate continuous random vector (X,Y) that forms a primitive input to a simulation. Here, the phrase "model a bivariate continuous random vector (X,Y)" is meant in the statistical sense, that is, we wish to choose a bivariate distribution that approximates the (true) unknown distribution $H(x,y) = \Pr(X \leq x, Y \leq y)$ of (X,Y). The traditional way of accomplishing such approximation is to collect data $(X_i, Y_i), i = 1, 2, \ldots, n$, choose a bivariate parametric family, e.g., the bivariate gamma (Kotz et al. 1997, Chapter 48), and then "fit a distribution" from the chosen family using one of various available methods (Casella and Berger 2002,

Chapter 7). While such a method is reasonable, it is limiting from the standpoint of distribution-family choice. For instance, if the marginal data reveal that X should have an exponential distribution and Y should have a normal distribution, the modeler is forced to search for a "named" bivariate family having the exponential and normal families as marginal distributions (or worse, model X and Y as independent random variables). This issue becomes more and more pronounced in higher dimensions, where the number of existing predefined multivariate distributions, especially with different marginal df families, is small.

The above issue leads to a natural question: is there a systematic and flexible way of modeling dependence in a random vector while preserving chosen marginal structures? The answer to this question might (implicitly) lie in a powerful representation theorem due to Sklar (Nelsen 2007, Section 2.3). To illustrate Sklar's theorem, suppose the bivariate distribution $H(x,y) = \Pr(X \le x, Y \le y)$ is continuous with marginal cdfs $F_X(x) = H(x,\infty)$ and $F_Y(y) = H(\infty,y)$. We see that the random variables $F_X(X)$ and $F_Y(Y)$ are distributed as U$[0,1]$ since $\Pr(F_X(X) \le x) = \Pr(X \le F_X^{-1}(x)) = x$ and $\Pr(F_Y(Y) \le y) = \Pr(Y \le F_Y^{-1}(y)) = y$. The analogous normalization in the bivariate space is provided by Sklar's theorem (Nelsen 2007) which states that if the joint distribution function $H(x,y)$ is continuous, then it has the unique representation

$$H(x,y) = C(F_X(x), F_Y(y)), \tag{2}$$

where the function $C : [0,1] \times [0,1] \to [0,1]$ is called the *copula* associated with (X,Y).

The implication of the representation in (2) is that the joint cdf of (X,Y) at any point can be expressed purely as a function of the marginal cdfs F_X and F_Y through the copula C. For example, when X and Y are independent, we can write

$$P(X \le x, Y \le y) = F_X(x)F_Y(y) = C(F_X(x), F_Y(y)), \text{ where } C(u,v) = uv, (u,v) \in [0,1] \times [0,1].$$

In this case, we say that the function $C(u,v) = uv, (u,v) \in [0,1] \times [0,1]$ is the copula associated with *any* two independent continuous random variables X and Y. Similarly, when (X,Y) has a bivariate normal distribution with covariance matrix Σ, we can write

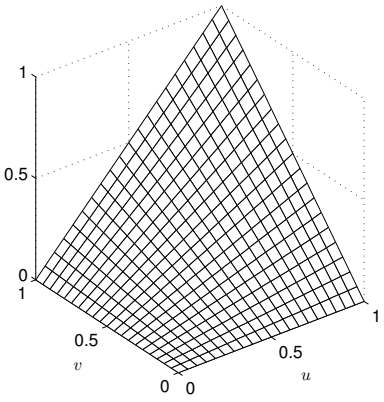
$$P(X \le x, Y \le y) := \Phi_\Sigma(x,y) = \Phi_\Sigma\left(\Phi^{-1}(\Phi(x)), \Phi^{-1}(\Phi(y)\right) := C_G(\Phi(x), \Phi(y))$$

giving the 2-dimensional Gaussian copula $C_G(u,v) = \Phi_\Sigma(\Phi^{-1}(u), \Phi^{-1}(v)), (u,v) \in [0,1] \times [0,1]$. (The independence copula and a Gaussian copula are illustrated in Figure 4.)

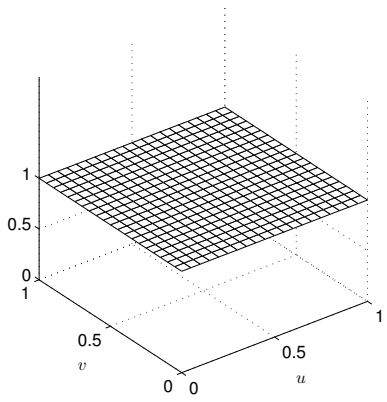
Remark 1 As we shall see in the next section, the representation for the df of a random vector (X,Y) that is discontinuous takes a form that is identical to (2). The uniqueness of the copula, however, is no longer guaranteed. In other words, for the discrete (or mixed) (X,Y) context, we can in principle find several copulas that satisfy the representation (2).

Remark 2 As noted in Resnick (1987), the standardization of the marginal dfs of H to uniform dfs on $[0,1]$ is arbitrary.

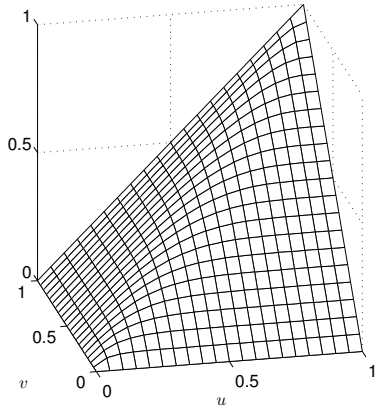
The representation characterized through Sklar's theorem is a powerful tool for modeling dependence in simulation input. To see this, let us go back to the example of a bivariate continuous random vector (X,Y) that forms the input to a simulation. With Sklar's theorem, we know that the bivariate df $H(x,y)$ of (X,Y) can be written as $H(x,y) = C(F_X(x), F_Y(y))$, where $C(u,v) : [0,1] \times [0,1] \to [0,1]$ is a copula satisfying certain regularity conditions. A model for $H(x,y)$ is thus completely specified upon *individually* modeling the three ingredients that comprise the bivariate distribution function $H(x,y)$: the marginal distribution function F_X, the marginal distribution F_Y, and the copula $C(u,v)$. Such individual modeling is convenient in that it allows the modeler to "mix and match" different choices of continuous marginal distributions with copula functions that encode different dependence structures. For instance, a modeler might choose a gamma(1,2) (Johnson et al. 1995, Chapter 17) distribution to model F_X, an exponential(1) (Johnson et al. 1995, Chapter 19) distribution to model F_Y, and a Gaussian copula to model the dependence between X and Y. Such flexibility is in contrast to the more traditional way of modeling $H(x,y)$ by a member



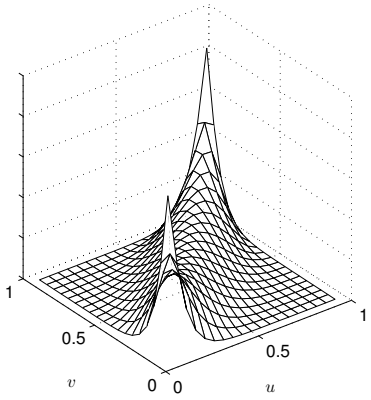
(a) Independence Copula



(b) Density of Independence Copula



(c) Gaussian Copula



(d) Density of Gaussian Copula

Figure 4: The independence copula and the Gaussian copula

of a prespecified parametric family of bivariate distributions having gamma and exponential marginal distributions. Such a specific bivariate family may not be readily available.

A second attractive feature that is implicit in the representation $H(x,y) = C(F(x),F(y))$ is the *invariance property* associated with the copula C. Specifically, if $\psi_1 : \mathbb{R} \to \mathbb{R}$ and $\psi_2 : \mathbb{R} \to \mathbb{R}$ are real-valued strictly increasing (or strictly decreasing) functions, then the random vector $(\psi_1(X), \psi_2(Y))$ has the marginal distributions $F_{\psi_1(X)}(x) = P(\psi_1(X) \leq x) = F_X(\psi_1^{-1}(x)), F_{\psi_2(Y)}(x) = P(\psi_2(Y) \leq y) = F_Y(\psi_2^{-1}(y))$, and the df

$$\begin{aligned} P(\psi_1(X) \leq x, \psi_2(Y) \leq y) = H(\psi_1^{-1}(x), \psi^{-1}(y)) &= C(F_X(\psi_1^{-1}(x)), F_Y(\psi_2^{-1}(y))) \\ &= C(F_{\psi_1(X)}(x), F_{\psi_2(Y)}(x)). \end{aligned} \tag{3}$$

We see from (3) that under the strictly increasing (or strictly decreasing) transformations (ψ_1, ψ_2), the random vector (X_1, X_2) retains its dependence structure in the sense that the transformed random vector $(\psi_1(X), \psi_2(Y))$ has the same copula as the original random vector (X,Y). As we shall see later, such invariance under monotone transformations suggests that the relative rank $(F_X(X), F_Y(Y))$ of the random vector (X,Y) contains complete information on the dependence structure — a feature that can be exploited for fitting copulas, since the sample relative ranks can be readily estimated from available data.

A third implicit advantage is afforded by Sklar's representation. Since the copula C associated with a random vector (X,Y) is essentially a form of standardization of the dependence structure implicit in (X,Y), the copula provides a coherent way to compare the extent of dependence of random vectors having different marginal distributions. So, for example, the dependence implicit in a bivariate exponential having specific parameter values can be compared against the extent of dependence implicit in a bivariate normal having a specified covariance matrix, through their respective copulas. Admittedly, such comparison is difficult in dimensions higher than two. In fact, even in two dimensions, interpreting the contours of a copula for comparison against another can be an arduous task. There is reason to believe, however, that any reasonable measure of comparison of dependence structures in high dimensions will face similar challenges of interpretation.

3 COPULAS: FORMAL DEFINITION AND SOME BASIC PROPERTIES

In this section, we formally define copulas, state Sklar's Theorem, and discuss some important properties. Depending on the reader's objective, the details of this section can be safely omitted, but only after internalizing Definition 1 and Theorem 1.

We start with the definition of a copula.

Definition 1 A real-valued function $C:[0,1]^d \to [0,1]$ is called a d-dimensional *copula* if it is a d-dimensional df with uniform marginals, that is, it satisfies the following three properties.

- P.1 $C(u_1,u_2,\ldots,u_d)$ is non-decreasing in each of its arguments $u_1,u_2,\ldots,u_d$.
- P.2 $C(1,1,\ldots,1,u_i,1,\ldots,1) = u_i$ for $i = 1,2,\ldots,d$.
- P.3 Suppose $a_j \le b_j, j = 1,2,\ldots,d$ and $u_{j,1} = a_j, u_{j,2} = b_j$. The function $C(u_1,u_2,\ldots,u_d)$ satisfies the *rectangle inequality*

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+i_2+\cdots+i_d} C(u_{1,i_1},\ldots,u_{d,i_d}) \ge 0.$$

The property P.1 is the usual restriction imposed on one-dimensional dfs. The property P.2 stipulates that the ith marginal function, obtained by setting all arguments except the ith to one, corresponds to the U$[0,1]$ distribution. The property P.3 is a formal restatement that the measure $\mathrm{P}(U_1 \in [a_1,b_1],\ldots,U_d \in [a_d,b_d]) \ge 0$ associated with rectangular sets $\Pi_{j=1}^{d}[a_j,b_j]$ is non-negative. Apart from P.1, P.2, and P.3, the definition of the copula endows it with the other natural property that if $1 \le j_1 \le j_2 \le \cdots \le j_k \le d$ are the labels of k arguments, then the function $C(1,\ldots,1,j_1,1,\ldots,1,j_2,1,\ldots,1,j_k,\ldots)$ obtained by setting all arguments other than those at positions $j_1,j_2,\ldots,j_k$ to one constitutes a k-dimensional copula.

We are now ready to formally state Sklar's theorem (Nelsen 2007, Section 2.3).

Theorem 1 (**Sklar's Theorem**) Let H be a d-dimensional cdf with corresponding (univariate) marginal dfs $H_1,H_2,\ldots,H_d$. Then, there exists a d-dimensional copula C such that

$$H(\mathbf{x}) = C(H_1(x_1),H_2(x_2),\ldots,H_d(x_d)).$$

The copula C is unique if H is a continuous df. Otherwise, C is unique on the set $\mathscr{D}_1 \times \mathscr{D}_2 \times \cdots \times \mathscr{D}_d$, where $\mathscr{D}_i$ is the range of $H_i, i = 1,2,\ldots,d$.

From the definition of the df H, it is easy to see that a copula C satisfies

$$C(\mathbf{u}) = H(H_1^{-1}(u_1),H_2^{-1}(u_2),\ldots,H_d^{-1}(u_d)). \tag{4}$$

Sklar's theorem notes that the above copula is unique when the df H is continuous. Using the representation of the copula in (4), and assuming that the df H has a joint density $h := (h_1,h_2,\ldots,h_d)$, the density function $c:[0,1]^d \to [0,1]$ associated with the copula C is given as

$$c(u_1,u_2,\ldots,u_d) = \frac{h(H_1^{-1}(u_1),\ldots,H_d-1(u_d))}{h_1(H_1^{-1}(u_1))\cdots h_d(H_d^{-1}(u_d))}. \tag{5}$$

Being distribution functions, copulas are often difficult to interpret visually; their densities, when existent, are more amenable to visual interpretation. Contour plots of some examples of copula densities appear in the ensuing section.

An important property that makes the copula an attractive measure of dependence is its invariance under monotone transformations. This property is formally stated in the following theorem.

Theorem 2 (**Invariance Under Monotone Transforms**) Suppose the random vector $X := (X_1, X_2, \ldots, X_d)$ has the df $H := (H_1, H_2, \ldots, H_d)$ and the associated d-dimensional copula $C : [0,1]^d \to [0,1]$, that is,

$$H(\mathbf{x}) = C(H_1(x_1), H_2(x_2), \ldots, H_d(x_d)).$$

Let $\psi_i : \mathbb{R} \to \mathbb{R}, i = 1, 2, \ldots, d$ be strictly increasing functions. Then, the copula associated with the random vector $(\psi_1(X_1), \psi_2(X_2), \ldots, \psi_d(X_d))$ remains C, that is,

$$\mathrm{P}(\psi_1(X_1) \le x_1, \psi_2(X_2) \le x_2, \cdots, \psi_d(X_d) \le x_d) = C(H_{\psi_1(X_1)}(x_1), H_{\psi_2(X_2)}(x_2), \ldots, H_{\psi_d(X_d)}(x_d)),$$

where $H_{\psi_i(X_i)}(x) = H_1(\psi_1^{-1}(x)$ is the df of $\psi_i(X_i)$.

Theorem 2, in a sense, points to the copula as the correct measure of dependence. Loosely speaking, the copula of a random vector is its df with arguments measured on the "marginal scale"; Theorem 2 simply asserts that the relative ranks (as measured on the "marginal scale") of random variables remain unchanged under strictly monotone transformations.

The maximum and minimum possible (Pearson) correlation between any two (real-valued) random variables having the respective dfs F_1 and F_2 are given by $\mathrm{Corr}(F_1^{-1}(U), F_2^{-1}(U))$ and $\mathrm{Corr}(F_1^{-1}(U), F_2^{-1}(1-U))$ respectively, where $U \sim U[0,1]$ (Whitt 1976). This result is well-known among simulationists in a more practical sense. Namely, in order to induce the maximum and minimum amount of correlation between two random variates X_1 and X_2 having dfs F_1 and F_2, generate (X_1, X_2) (using the cdf-inverse method) with common random numbers or antithetic variates respectively, that is, generate $X_1 \leftarrow F_1^{-1}(U), X_2 \leftarrow F_2^{-1}(U)$ for maximum correlation, and generate $X_1 \leftarrow F_1^{-1}(U), X_2 \leftarrow F_2^{-1}(1-U)$ for minimum correlation. A corresponding result holds in the context of copulas and is known as the Fréchet-Hoeffding bound.

Theorem 3 (**Fréchet-Hoeffding bound**) The d-dimensional copula C satisfies

$$\max\{\sum_{i=1}^{d} u_i + 1 - d, 0\} \le C(u_1, u_2, \ldots, u_d) \le \min\{u_1, u_2, \ldots, u_d\}.$$

The upper bound comes from the simple calculation that $C(u_1, u_2) = P(U_1 \le u_1, U_2 \le u_2) \le P(U_1 \le u_1, U_1 \le u_2) = \min(u_1, u_2)$. A similar calculation holds for the lower bound for $d = 2$ by setting $U_2 = 1 - U_1$; the lower bound calculation for dimensions higher than 2 is more involved.

Remark 3 The Fréchet-Hoeffding bounds are attained (for $d = 2$) by many of the popular families that will be listed in the subsequent section. The Fréchet-Hoeffding lower bound cannot be attained for $d > 2$ since the lower bound is not a copula.

4 POPULAR COPULA FAMILIES

Recall the setting of modeling simulation input. The premise is that there exists a primitive input random vector $(X_1, X_2, \ldots, X_d)$ whose df $H(\mathbf{x})$ we wish to approximate (or "model") with an appropriate choice. To facilitate such choice, we relied on Sklar's theorem to write

$$H(\mathbf{x}) = C(H_1(x_1), H_2(x_2), \ldots, H_d(x_d)),$$

where the H_is are the marginal dfs and C is a d-dimensional copula that encodes the dependence structure in standardized form. The advantage of Sklar's decomposition is that the marginal dfs can be chosen

individually, and based on the available marginal data on each dimension. The big question is how should one model the dependence that is inherent in $(X_1, X_2, \ldots, X_d)$, which essentially amounts to choosing the copula C that connects the chosen marginal dfs. Since the copula is standardized, that is, it is a df having uniform marginals, a discussion about its choice can happen without reference to the chosen marginal dfs.

In this section, in an attempt to guide the choice of a copula, we discuss common copula families. Just as is the case with choosing univariate marginal distributions, choosing a copula family is at least in part based on subjective measures. The appropriateness of a (parameterized) copula family depends on the ability of the family to model different shapes, the ease with which parameters are chosen (or "fitted" using data), and the ease with which random variates can be generated from the family. In what follows, we list a number of common copula families along with a discussion on their general shapes and structural properties. For easier interpretation, we depict the density functions of the copulas we discuss (rather than the copulas themselves). The question of estimation is taken up in later sections.

4.1 Independence Copula

As the name suggests, the independence copula amounts to specifying no dependence structure between the random variables $X_1, X_2, \ldots, X_d$, that is, the independence copula is the copula associated with independent random variables $X_1, X_2, \ldots, X_d$. For this reason, the independence copula has limited modeling value, although important for understanding as one of the simplest copulas. Since $X_1, X_2, \ldots, X_d$ are independent, we can write

$$H(\mathbf{x}) = \prod_{i=1}^{d} H_i(x_i) := C_I(H_1(x_1), H_2(x_2), \ldots, H_d(x_d)),$$

leading to the independence copula $C_I(\mathbf{u}) = \prod_{i=1}^{d} u_i, \mathbf{u} \in [0,1]^d$ and corresponding constant density function $c_I(\mathbf{u}) = 1$.

4.2 Implicit Copulas

An easy way to induce dependence between random variables having (any specified) marginal dfs is by incorporating copulas derived from "named" multivariate distributions. For example, the Gaussian copula is the copula associated with the multivariate Gaussian df. Specifically, suppose

$$\phi_\rho(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\rho|}} \exp\left(-\frac{1}{2}\mathbf{x}^T \rho^{-1} \mathbf{x}\right)$$

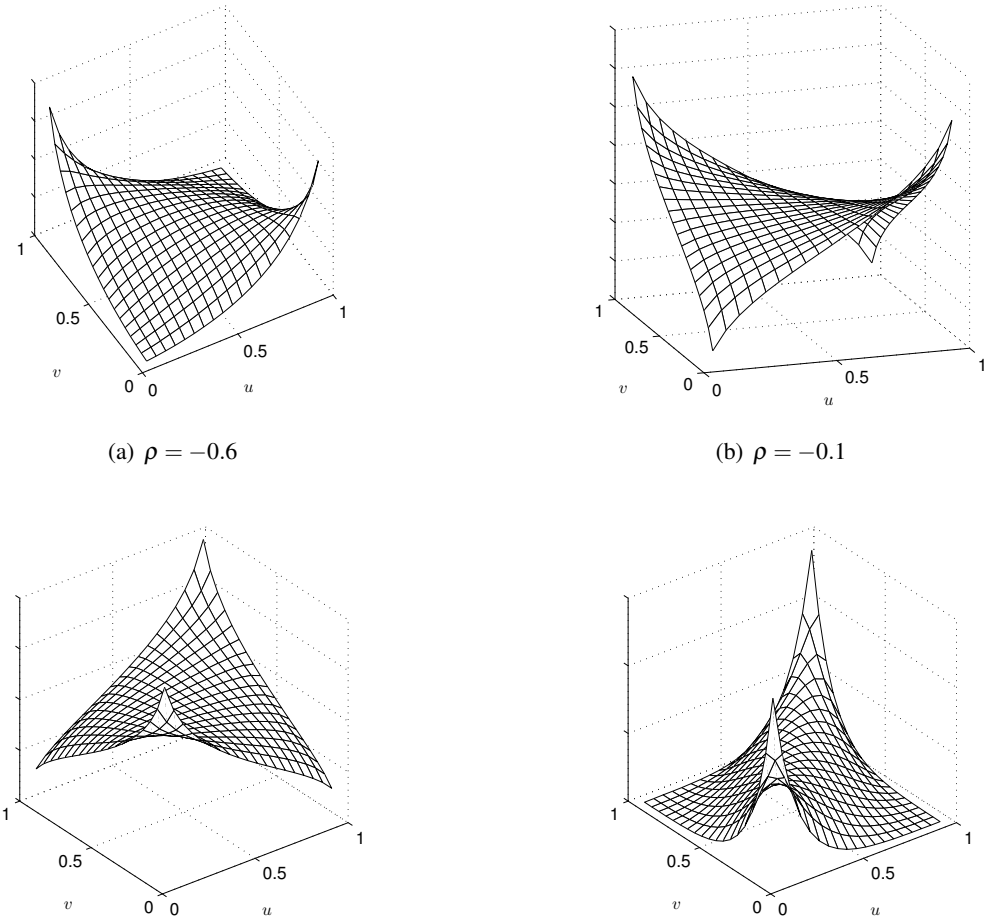
is the d-dimensional Gaussian density having standard normal marginals, then the corresponding df $\Phi_\rho(\mathbf{x})$ can be decomposed as

$$\begin{aligned} \Phi_\rho(\mathbf{x}) &= \Phi_\rho(\Phi^{-1}(\Phi(x_1)), \Phi^{-1}(\Phi(x_2)), \ldots, \Phi^{-1}(\Phi(x_d))) = C_G(\Phi(x_1), \Phi(x_2), \ldots, \Phi_d(x_d)); \\ C_G(\mathbf{u}) &= \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d)), \quad \mathbf{u} \in [0,1]^d \end{aligned} \tag{6}$$

where the d-dimensional function $C_G(\mathbf{u})$ is called the Gaussian copula. The density associated with the Gaussian copula can be calculated explicitly using the formula in (5).

Remark 4 A Gaussian distribution having the correlation matrix ρ, mean vector μ, and marginal variances $\sigma_i^2, i = 1, 2, \ldots, d$ has the same copula $C_G(\mathbf{u})$ as a Gaussian with standard Gaussian marginal dfs and correlation matrix ρ. This is evident through the usual standardizing argument applied to the general Gaussian random vector, or by directly applying Theorem 2.

Owing to Remark 4, notice that the Gaussian copula has no location or scale parameters; it is in this sense that the Gaussian copula (or any other copula) is purely a measure of the extent of the dependence that exists between the random variables constituting the random vector being modeled. Figure 5 depicts the bivariate Gaussian copula density functions associated with a few different parameter settings.



(a) $\rho = -0.6$

(b) $\rho = -0.1$

(c) $\rho = 0.3$

(d) $\rho = 0.8$

Figure 5: Gaussian Copula Densities

More generally, and following the arguments leading to (6), the copula C_H associated with any continuous df $H(\mathbf{x})$ is given as

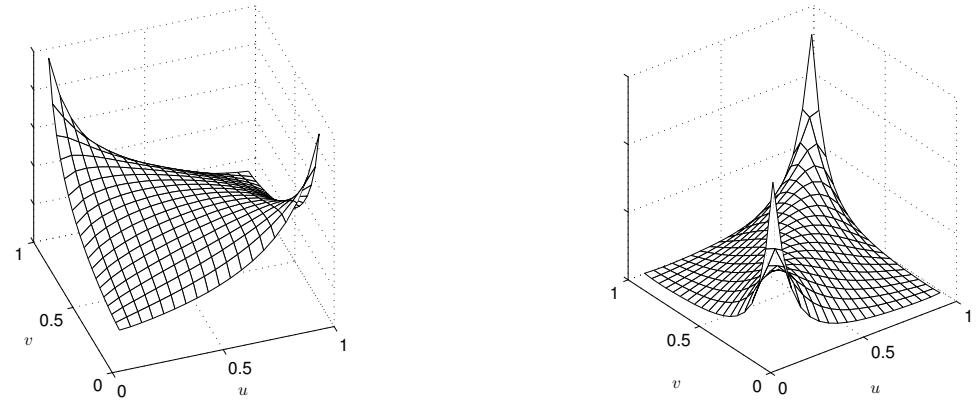
$$C_H(\mathbf{u}) = H(H_1^{-1}(u_1), H_2^{-1}(u_2), \ldots, H_d^{-1}(u_d)), \tag{7}$$

where $H_i, i = 1,2,\ldots,d$ is the ith marginal associated with H. Using (7), we see that the Student's T copula is given as

$$C_T(\mathbf{u}) = t_{\nu,\rho}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)),$$

where ρ is a correlation matrix, t_ν is the df of the one-dimensional Student's T distribution, and $t_{\nu,\rho}$ is the df of the multivariate Student's T distribution (Kotz et al. 1997). Since a Student's T random vector $(T_1, T_2, \ldots, T_d)$ with ν degrees of freedom has the same distribution as $(\frac{Z_1}{\|Z/\nu\|_2}, \frac{Z_1}{\|Z/\nu\|_2}, \cdots, \frac{Z_1}{\|Z/\nu\|_2})$, where $(Z_1, Z_2, \ldots, Z_d)$ has the multivariate Gaussian df with standard normal marginals, the Student's T copula has $d \times (d-1)/2$ parameters. Figure 6 depicts the density function associated with the Student's T copula for various parameter settings and $d = 2$.

A copula that appears to be very useful from the standpoint of modeling flexibility is the generalized gamma copula C_Γ, derived from one of the numerous multivariate gamma distribution definitions (Viraswami 1991, Kotz et al. 1997). We do not go into further detail here but note that since all of the available



(a) $\rho = -0.35, \nu = 7$ (b) $\rho = 0.65, \nu = 3$

Figure 6: Student's T Copula Densities

forms of the multivariate gamma are continuous, the corresponding copula density follows along the lines outlined for the Gaussian and the Student's T copulas.

Copulas associated with named families of multivariate distributions provide for a convenient way of incorporating dependence modeling. We emphasize that the copula associated with any continuous multivariate family, e.g., Gaussian or the generalized gamma, can be used alongside any set of chosen marginal distribution functions. Thus, a modeler choosing marginal dfs $F_1, F_2, \ldots . F_d$ and (say) the generalized gamma copula C_Γ to model dependence has implicitly chosen the multivariate df $C_\Gamma(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$ as the model for her simulation input.

4.3 Archimedean Copulas

In contrast to copulas derived from other "named" multivariate distributions, copulas can be defined directly and in such a way that the regularity properties listed in Definition 1 are satisfied. The Archimedean copulas constitute one such family, and are defined in a way that facilitates easy construction of new copulas. Formally, a 2-dimensional copula is called an Archimedean copula if it admits the representation

$$C_A(\mathbf{u}) = \eta^{-1}(\eta(u_1) + \eta(u_2)), \tag{8}$$

where the *generator function* $\eta : [0,1] \to [0,\infty)$ is a continuous strictly decreasing convex function (Boyd 2004) satisfying $\eta(1) = 0$ and

$$\eta^{-1}(t) = \begin{cases} \eta^{-1}(t) & \text{if } 0 \le t \le \eta(0); \\ 0 & \text{if } \eta(0) \le t \le \infty. \end{cases} \tag{9}$$

(The function $C_A(\mathbf{u})$ is a copula if and only if $\eta : [0,1] \to [0,\infty]$ is continuous, strictly decreasing, and convex, with $\eta(1) = 0$ (Nelsen 2007, pp. 91,92).) Archimedean copulas form a metafamily in the sense that they are defined to facilitate easy construction of new copulas by specifying different generator functions $\eta(t)$, which are themselves usually parameterized with one of more parameters.

Commonly considered Archimedean copulas include the two-parameter Clayton copula with the generator $\eta(t) = \theta^{-\delta}(t^{-\theta} - 1)^{\delta}, \theta \ge 0, \delta \ge 0$; the Frank copula with the generator function $\eta(t) = -\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right), \theta \in \mathbb{R}$; and the Gumbel copula with the generator function $\eta(t) = (-\ln t)^{\theta}, \theta \ge 1$. The Frank and Clayton copulas are comprehensive in the sense that they interpolate between the lower and upper Fréchet-Hoeffding bounds as the parameter θ tends to 0 and ∞.

The natural extension of the bivariate Archimedean copula to higher dimensions is

$$C_A(\mathbf{u}) = \eta^{-1}(\eta(u_1) + \eta(u_2) + \cdots + \eta(u_d)), \tag{10}$$

where, again, the *generator function* $\eta : [0,1] \to [0,\infty)$ is a continuous strictly decreasing convex function satisfying $\eta(1) = 0$. It so happens, however, that an additional condition is needed on the generator inverse function to guarantee that the d-dimensional function in (10) is indeed a copula. Specifically, the generator inverse $\eta^{-1} : [0,\infty] \to [0,1]$ should be *completely monotonic*, that is, $(-1)^k \frac{d^k}{dt^k}\eta^{-1}(t) \geq 0, k \in \mathbb{N}, t \in [0,\infty]$ in order that the function C_A in (10) is a copula (Kimberling 1974). The generator functions for Gumbel, Clayton, and Frank copulas are completely monotonic. Archimedean copulas have seen limited success in dimensions greater than 2 or 3, primarily due to the fact that the structure imposed through (10) has been recognized to be restrictive (McNeil et al. 2005).

5 COPULA ESTIMATION

In this section, we consider the question of fitting a df, that is, estimating a specific distribution to model the primitive input vector $\mathbf{X}$ to a simulation. For concreteness, assume that the data vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, which are identically distributed copies of the random vector $\mathbf{X}$, are available to us. Following the framework we have proposed in this paper, we assume that the model df takes the form

$$\tilde{H}(\mathbf{x}) = C_\theta(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)),$$

where C_θ is a chosen (parametric) copula family, θ is the parameter vector, and $F_i, i = 1,2,\ldots,d$ are the chosen marginal dfs which may or may not have their own parameters. Fitting the model $\tilde{H}$ then means estimating the copula parameters θ and the marginal dfs.

Remark 5 Just as in the univariate case, one can obtain a non-parametric estimate of the joint df in the usual manner

$$\tilde{H}(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}(\mathbf{X}_i \leq \mathbf{x}), \tag{11}$$

or through the kernel method

$$\tilde{H}(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{K}(\frac{\mathbf{x} - \mathbf{X}_i}{h}) \tag{12}$$

with a d-dimensional kernel K satisfying $\mathbb{K}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} K$. Both of these estimates lead to corresponding notions of empirical copulas.

5.1 Method of Moments

A straightforward way of estimating the parameters θ of the chosen copula C_θ is by matching all pairwise rank correlation(s) associated with the distribution $C_\theta(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$ to the empirical rank correlation obtained from the observed data. The suggested rank correlation is usually either Spearman's rank correlation, denoted here as ρ_S, or Kendall's rank correlation, denoted here as ρ_τ. (See (Stuart et al. 1996) for a full treatment of rank correlations.) The sample estimators of the Spearman's and Kendall's rank correlations between the ith and jth components of $\mathbf{X}$ are given by

$$\hat{\rho}_S(i,j) = \frac{12}{n(n^2-1)}\sum_{t=1}^{n}\left(\overline{F}_{i,n}(X_{t,i}) - \frac{1}{2}(n+1)\right)\left(\overline{F}_{j,n}(X_{t,j}) - \frac{1}{2}(n+1)\right)$$

and

$$\hat{\rho}_\tau(i,j) = \binom{n}{2}^{-1} \sum_{1 \leq t < s \leq n} \text{sign}((X_{t,i} - X_{s,i})(X_{t,j} - X_{s,j})),$$

where $\overline{F}_{i,n}$ is the empirical df of the ith marginal distribution.

Given a sample estimate of the (Spearman's or Kendall's) rank correlation, a parameter estimate for θ can be obtained by equating the theoretical value of the rank correlation for the assumed copula C_θ with the sample estimate, and then solving for a parameter estimate. For example, if the assumed copula is a Gaussian copula, the theoretical value of the (i,j)th Spearman's correlation is given as $\frac{6}{\pi}\arcsin\frac{\rho(i,j)}{2}$, where $\rho(i,j)$ is the (i,j)th element of the matrix ρ used in the Gaussian copula. (For the Gaussian copula $\rho := \theta$.) Equating this to the sample Spearman's rank correlation yields the estimate $\hat{\rho}(i,j) = 2\sin(\hat{\rho}_S * \frac{\pi}{6})$.

A similar calculation for the Student's T copula but using Kendall's rank correlation yields the estimate $\hat{\rho}(i,j) = \sin(\hat{\rho}_\tau \frac{\pi}{2})$. Closed-form solutions for the parameter estimates (obtained through the method of moments and using Kendall's rank correlation) for various bivariate Archimedean copulas can be obtained through Table 5.5 in McNeil et al. (2005).

Remark 6 In the above treatment for estimating parameters, we could have used the more traditional Pearson (linear) correlation instead of the rank correlation. We emphasize, however, that the Pearson correlation is generally considered inferior to the rank correlation as a measure of dependence. Unlike rank correlation, Pearson correlation is generally not invariant with respect to strictly monotone transformations. (Pearson correlation is invariant under linear transformations.) Pearson correlation is a natural measure of dependence only in the context of elliptical distributions.

Four comments regarding the estimation of parameters using the method of moments are noteworthy.

(i) Parameter estimates obtained by equating sample rank correlation estimates with the theoretical expressions for the rank correlation do not always yield closed-form solutions. For example, when using Spearman's rank correlation, the parameter estimate for the Student's T copula cannot be solved in closed-form. In such cases, the parameter estimates are obtained numerically, using one of numerous available techniques (Ortega and Rheinboldt 1970).

(ii) When the chosen marginal distributions are continuous, the parameter estimates obtained using the method of moments do not depend on the marginal distributions. It is in this sense that inference about the copula parameters in such a case is said to be "margin free." When the marginal dfs are not continuous (as is often the case), the parameter estimates are a function of the marginal dfs, implying that the marginal dfs need to be estimated before applying the method of moments. See Avramidis et al. (2009) for more on this.

(iii) The method of moments equates all pairwise sample rank correlations with the theoretical values, yielding a nonlinear system of $d \times (d-1)/2$ equations in $d \times (d-1)/2$. Clearly, if there are not enough number of parameters in the chosen copula, that is, if there are not enough degrees of freedom, the method of moments will not apply and an alternate method such as maximum likelihood estimation should be undertaken.

(iv) Inference on the parameters obtained using the method of moments is now well-understood (Genest et al. 1995).

5.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is the more general method of estimating parameters for the parametric copula C_θ). Assuming that the marginal dfs have been estimated, the MLE problem becomes

$$\max_\theta \sum_{t=1}^{n} \ln c_\theta(\overline{F}_1(X_{t,1}), \overline{F}_2(X_{t,2}), \ldots, \overline{F}_d(X_{t,d})), \tag{13}$$

where c_θ is the density of the copula C_θ. For example, for the Gaussian copula, the problem in (13) becomes

$$\max_\rho \sum_{t=1}^{n} \ln \phi_\rho(\Phi^{-1}(\overline{F}_1(X_{t,1})), \Phi^{-1}(\overline{F}_2(X_{t,2})), \ldots, \Phi^{-1}(\overline{F}_d(X_{t,d}))) - \sum_{t=1}^{n}\sum_{j=1}^{d} \ln \phi(\Phi^{-1}(\overline{F}(X_{t,j}))), \tag{14}$$

where ρ is a correlation matrix, and ϕ_ρ is the multivariate Gaussian density with standard Gaussian marginal dfs and correlation matrix ρ. (The second term in (14) is a constant that will not play a role in the optimization.)

The solution of the problem (13) is almost never in closed-form implying that a suitable numerical optimization technique (Boyd 2004) needs to be invoked. Inference on the solution obtained through MLE is now understood; see, for instance, Joe (1997) and Genest and Rivest (1993).

6 CONCLUDING REMARKS

When representing a physical system of interest using a simulation model, it is crucial to ensure that the dependence structure inherent in the primitive inputs to the simulation is faithfully represented. Ignoring such dependence, as is often done for the sake of convenience, can result in substantial modeling errors and consequent poor decision-making. Copulas constitute a flexible and convenient way to model dependent inputs because they standardize dependence structures and provide a mechanism to break down multivariate distribution functions into their marginal and dependent elements. Such a breakdown and standardization allows the modeler to individually select and estimate the marginal and dependent elements of the input model, thereby allowing for a more faithful representation. Estimation, inference, and random variate generation in the copula setting are now well-understood. Interestingly, there appears to be only a limited adoption of copulas into discrete-event simulation software packages at the time of this writing.

REFERENCES

Avramidis, A. N., N. Channouf, and P. L'Ecuyer. 2009. "Efficient Correlation Matching for normal-copula dependence when univarite marginals are discrete". *INFORMS Journal on Computing* Winter 2009 (21): 88–106.

Boyd, S. 2004. *Convex Optimization*. Cambridge, U.K.: Cambridge University Press.

Casella, G., and R. L. Berger. 2002. *Statistical Inference*. Second ed. Pacific Grove, CA: Duxbury.

Genest, C., K. Ghoudi, and L. P. Rivest. 1995. "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions". *Biometrika* 82:543–552.

Genest, C., and L. P. Rivest. 1993. "Statistical inference procedures for bivariate Archimedean copulas". *Journal of the American Statistical Association* 88:1034–1043.

Glasserman, P., and Y. Wang. 1998. "Leadtime-Inventory Trade-Offs in Assemble-To-Order Systems". *Operations Research* 46 (6): 858–871.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1995. *Continuous Univariate Distributions*, Volume 2. New York, NY.: John Wiley & Sons, Inc.

Kimberling, C. H. 1974. "A probabilistic interpretation of complete monotonicity". *Aequationes Mathematicae* 10:152–164.

Kotz, S., N. Balakrishnan, and N. L. Johnson. 1997. *Continuous Multivariate Distributions*. Second ed, Volume I: Models and Applications. New York, NY.: John Wiley & Sons, Inc.

Law, A. M. 2007. *Simulation Modeling and Analysis*. New York, NY.: McGraw-Hill.

Leemis, L. 2003. "Input Modeling". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. Sánchez, D. Ferrin, and D. Morrice, 14–24: Institute of Electrical and Electronics Engineers: Piscataway, New Jersey.

McNeil, A. J., R. Frey, and P. Embrechts. 2005. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton, NJ: Princeton University Press.

Nelsen, R. B. 2007. *An Introduction to Copulas*. New York, NY.: Springer.

Nelson, B. L. 1987. "A perspective on variance reduction in dynamic simulation experiments". *Communications in Statistics* B16:385–426.

Ortega, J. M., and W. C. Rheinboldt. 1970. *Iterative Solution of Nonlinear Equations in Several Variables*. New York, NY.: Academic Press.

Resnick, S. 1987. *Extreme Values, Regular Variation, and Point Processes*. New York, NY: Springer.

Song, J.-S., and P. Zipkin. 2003. "Supply chain operations: assemble-to-order systems". In *Supply Chain Management: Design, Coordination and Operation*, edited by S. Graves and T. D. Kok, Volume 11 of *Handbooks in Operations Research and Management Science*, Chapter 11. Amsterdam: Elsevier.

Stuart, A., K. Ord, and S. Arnold. 1996. *Kendall's Adavanced Theory of Statistics: Classical Inference and the Linear Model*, Volume 2A. New York, NY.: Oxford University Press Inc.

Viraswami, K. 1991. *On Multivariate Gamma Distributions*. Ph. D. thesis, Department of Mathematical Statistics, McGill University, Montreal, Quebec.

Whitt, W. 1976. "Bivariate Distributions with Given Marginals". *The Annals of Statistics* 4 (6): 1280–1289.

Xu, S. H. 1999. "Structural Analysis of a Queueing System with Multiclasses of Correlated Arrivals and Blocking". *Operations Research* 47 (2): 264–276.

AUTHOR BIOGRAPHIES

RAGHU PASUPATHY is an associate professor in the Department of Statistics at Purdue University. His research interests lie broadly in Monte Carlo methods with a specific focus on simulation optimization. He is a member of INFORMS, IIE, and ASA, and serves as an associate editor for Operations Research and INFORMS Journal on Computing. He is the Area Editor for the Simulation Desk at IIE Transactions. His email address is pasupath@purdue.edu and his web page is http://web.ics.purdue.edu/ pasupath.

KALYANI NAGARAJ is a visiting assistant professor in the Department of Statistics at Purdue University. Her research interests include Monte Carlo methodology, with a focus on simulation optimization and random variate generation. Kalyani was the runner-up in the 2014 INFORMS Computing Society student paper competition. Her email address is kalyanin@purdue.com.