UNDERSTANDING RELATIONSHIPS USING COPULAS*

Edward W. Frees† and Emiliano A. Valdez‡

ABSTRACT

This article introduces actuaries to the concept of "copulas," a tool for understanding relationships among multivariate outcomes. A copula is a function that links univariate marginals to their full multivariate distribution. Copulas were introduced in 1959 in the context of probabilistic metric spaces. The literature on the statistical properties and applications of copulas has been developing rapidly in recent years. This article explores some of these practical applications, including estimation of joint life mortality and multidecrement models. In addition, we describe basic properties of copulas, their relationships to measures of dependence, and several families of copulas that have appeared in the literature. An annotated bibliography provides a resource for researchers and practitioners who wish to continue their study of copulas. For those who wish to use copulas for statistical inference, we illustrate statistical inference procedures by using insurance company data on losses and expenses. For these data, we (1) show how to fit copulas and (2) describe their usefulness by pricing a reinsurance contract and estimating expenses for pre-specified losses.

1. INTRODUCTION

As emphasized in "General Principles of Actuarial Science" (Committee on Actuarial Principles 1997), actuaries strive to understand stochastic outcomes of financial security systems. Because these systems are generally complex, outcomes are measured in several dimensions. Describing relationships among different dimensions of an outcome is a basic actuarial technique for explaining the behavior of financial security systems to concerned business and public policy decision-makers. This article introduces the concept of a "copula" function as a tool for relating different dimensions of an outcome.

Understanding relationships among multivariate outcomes is a basic problem in statistical science; it is not specific to actuarial science nor is it new. In the late nineteenth century, Sir Francis Galton made a fundamental contribution to understanding multivariate relationships with his introduction of regression analysis. In one dataset described in his 1885 presidential address to the anthropological section of the British Association of the Advancement of Sciences, Galton linked the distribution of heights of adult children to the distribution of their parents' heights. Galton showed not only that each distribution was approximately normal but also that the joint distribution could be described as a bivariate normal. Thus, the conditional distribution of adult children's height, given the parents' height, could also be described by using a normal distribution. As a by-product of his analysis, Galton observed that "tall parents tend to have tall children although not as tall as the parents" (and vice versa for short children). From this, he incorrectly inferred that children would "regress to mediocrity" in subsequent generations, hence suggesting the term that has become known as *regression analysis*. See Seal (1967) and Stigler (1986) for additional accounts of the works of Galton and other early contributors to statistical science.

Regression analysis has developed into the most widely applied statistical methodology; see, for example, Frees (1996) for an introduction. It is an important component of multivariate analysis because it allows researchers to focus on the effects of explanatory variables. To illustrate, in the Galton dataset of family heights, regression allows the analyst to

*This paper was originally presented at the 32nd Actuarial Research Conference, held August 6–8, 1997, at the University of Calgary, Calgary, Alberta, Canada.

†Edward W. (Jed) Frees, F.S.A., Ph.D., is Time Insurance Professor of Actuarial Science, School of Business, University of Wisconsin–Madison, 975 University Avenue, Madison, Wisconsin 53706.

‡Emiliano A. Valdez, F.S.A., is a doctoral student at the Actuarial Science Risk Management and Insurance Department, School of Business, University of Wisconsin–Madison, 975 University Ave., Madison, Wisconsin 53706-1323.

describe the effect of parents' height on a child's adult height. Regression analysis also is widely applied in actuarial science; as evidence, it is a required educational component of the two main actuarial bodies in the U.S. and Canada, the Society of Actuaries and the Casualty Actuarial Society.

Although widely applicable, regression analysis is limited by the basic setup that requires the analyst to identify one dimension of the outcome as the primary measure of interest (the dependent variable) and other dimensions as supporting or ''explaining'' this variable (the independent variables). This article examines problems in which this relationship is not of primary interest; hence, we focus on the more basic problem of understanding the distribution of several outcomes, a *multivariate* distribution. For an example in actuarial science, when two lives are subject to failure, such as under a joint life insurance or annuity policy, we are concerned with joint distribution of lifetimes. As another example, when we simulate the distribution of a scenario that arises out of a financial security system, we need to understand the distribution of several variables interacting simultaneously, not in isolation of one another.

The normal distribution has long dominated the study of multivariate distributions. For example, leading references on multivariate analysis, such as Anderson (1958) and Johnson and Wichern (1988), focus exclusively on the multivariate normal and related distributions that can be derived from normal distributions, including multivariate extensions of Student's t- and Fisher's F-distributions. Multivariate normal distributions are appealing because the marginal distributions are also normal. For example, in the Galton dataset, the distribution of adult children's height and the distribution of parents' height are each approximately normal, in isolation of the other. Multivariate normal distributions are also appealing because the association between any two random outcomes can be fully described knowing only (1) the marginal distributions and (2) one additional parameter, the correlation coefficient.

More recent texts on multivariate analysis, such as Krzanowski (1988), have begun to recognize the need for examining alternatives to the normal distribution setup. This is certainly true for actuarial science applications such as for lifetime random variables (Bowers et al. 1997, Chap. 3) and long-tailed claims variables (Hogg and Klugman 1984), where the normal distribution does not provide an adequate approximation to many datasets. An extensive literature in statistics deals with nonnormal multivariate distributions; see, for example, Johnson and Kotz (1973) and Johnson, Kotz and Balakrishnan (1997). However, historically many multivariate distributions have been developed as immediate extensions of univariate distributions, examples being the bivariate Pareto, bivariate gamma, and so on. The drawbacks of these types of distributions are that (1) a different family is needed for each marginal distribution, (2) extensions to more than just the bivariate case are not clear, and (3) measures of association often appear in the marginal distributions.

A construction of multivariate distributions that does not suffer from these drawbacks is based on the *copula* function. To define a copula, begin as you might in a simulation study by considering p uniform (on the unit interval) random variables, $u_1, u_2, \ldots, u_p$. Here, p is the number of outcomes that you wish to understand. Unlike many simulation applications, we do not assume that $u_1, u_2, \ldots, u_p$ are independent; yet they may be related. This relationship is described through their joint distribution function

$$C(u_1, u_2, \ldots, u_p) = \text{Prob}(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_p \leq u_p).$$

Here, we call the function C a *copula*. Further, U is a (ex-ante) uniform random variable, whereas u is the corresponding (ex-post) realization. To complete the construction, we select arbitrary marginal distribution functions $F_1(x_1), F_2(x_2), \ldots, F_p(x_p)$. Then, the function

$$C[F_1(x_1), F_2(x_2), \ldots, F_p(x_p)] = F(x_1, x_2, \ldots, x_p) \quad (1.1)$$

defines a multivariate distribution function, evaluated at $x_1, x_2, \ldots, x_p$, with marginal distributions $F_1, F_2, \ldots, F_p$.

With copula construction in Equation (1.1), we select different marginals for each outcome. For example, suppose we are considering modeling male and female lifetimes for a joint-life annuity product. Then, with $p=2$, we might choose the Gompertz distribution to represent mortality at the older ages, yet with different parameters to reflect gender differences in mortality. As another example, in Section 4, we consider a bivariate outcome associated with the loss and the expense associated with administering a property and casualty claim. There, we could elect to use a lognormal distribution for expenses and a longer tail distribution, such as Pareto, for losses associated with the claim. The copula construction does not constrain the choice of marginal distributions.

In Section 2 we see that the copula method for understanding multivariate distributions has a relatively short history in the statistics literature; most of the statistical applications have arisen in the last ten years. However, copulas have been studied in the probability literature for about 40 years (Schweizer 1991), and thus several desirable properties of copulas are now widely known. To begin, it is easy to check from the construction in Equation (1.1) that F is a multivariate distribution function. Sklar (1959) established the converse. He showed that any multivariate distribution function F can be written in the form of Equation (1.1), that is, using a copula representation. Sklar also showed that if the marginal distributions are continuous, then there is a unique copula representation. In this sense copulas provide a unifying theme for our study of multivariate distributions. Sections 3 and 5 describe other desirable properties of copulas.

Given that copulas are fundamental building blocks for studying multivariate distributions, we now turn to the question of how to build a copula function for a problem at hand. Despite Sklar's result that a copula function always exists, Example 1.1 shows that it is *not* always convenient to identify the copula. Example 1.2 illustrates a useful way of building a copula, using the method of *compounding*. We describe this method of constructing copulas in detail in Section 3.1.

Example 1.1 Marshall-Olkin (1967) Exponential Shock Model

Suppose that we wish to model $p=2$ lifetimes that we suspect are subject to some common disaster, or "shock," that may induce a dependency between the lives. For simplicity, let us assume that Y_1 and Y_2 are two independent (underlying) lifetimes with distribution functions H_1 and H_2. We further assume there exists an independent exponential random variable Z with parameter λ that represents the time until common disaster. Both lives are subject to the same disaster, so that actual ages-at-death are represented by $X_1 = \min(Y_1, Z)$ and $X_2 = \min(Y_2, Z)$. Thus, the marginal distributions are

$$\begin{aligned}\mathrm{Prob}\big(X_j \le x_j\big) &= F_j(x_j) \\ &= 1 - \exp(-\lambda x_j)\big(1 - H_j(x_j)\big), \text{ for } j = 1, 2.\end{aligned}$$

Basic calculations show that the joint distribution is[1]

$$\begin{aligned}F(x_1, x_2) &= F_1(x_1) + F_2(x_2) - 1 \\ &\quad + \exp\big(\lambda \max(x_1, x_2)\big)\big(1 - F_1(x_1)\big)\big(1 - F_2(x_2)\big).\end{aligned}$$

This expression, although intuitively appealing, is not in the form of the copula construction (1.1) because the joint distribution function F is not a function of the marginals $F_1(x_1)$ and $F_2(x_2)$. For further discussions in the actuarial literature of this bivariate distribution, see Frees (1996) and Bowers et al. (1997, Sec. 9.6).

Example 1.2 Bivariate Pareto Model

Consider a claims random variable X that, given a risk classification parameter γ, can be modeled as an exponential distribution; that is,

$$\mathrm{Prob}\big(X \le x|\gamma\big) = 1 - e^{-\gamma x}.$$

As is well known in credibility theory (see, for example, Klugman et al. 1997), if γ has a gamma distribution, then the marginal distribution (over all risk classes) of X is Pareto. That is, if γ is gamma(α,λ), then

$$\begin{aligned}F(x) &= \mathrm{Prob}\big(X \le x\big) \\ &= \int \mathrm{Prob}\big(X \le x|\gamma\big)\, \frac{\alpha^{\lambda}}{\Gamma(\lambda)}\, \gamma^{\alpha-1}\, e^{-\lambda\gamma}\, d\gamma \\ &= 1 - \int e^{-\gamma x}\, \frac{\alpha^{\lambda}}{\Gamma(\lambda)}\, \gamma^{\alpha-1}\, e^{-\lambda\gamma}\, d\gamma \\ &= 1 - (1 + x/\lambda)^{-\alpha}, \qquad (1.2)\end{aligned}$$

a Pareto distribution.

Suppose, conditional on the risk class γ, that X_1 and X_2 are independent and identically distributed. Assuming that they come from the same risk class γ induces a dependency. The joint distribution is[2]

[1] $$\begin{aligned}F(x_1,x_2) &= \mathrm{Prob}\big(X_1 \le x_1, X_2 \le x_2\big) = 1 - \mathrm{Prob}\big(X_1 > x_1\big) - \mathrm{Prob}\big(X_2 > x_2\big) + \mathrm{Prob}\big(X_1 > x_1, X_2 > x_2\big) \\ &= 1 - \exp(-\lambda x_1)(1 - H_1(x_1)) - \exp(-\lambda x_2)(1 - H_2(x_2)) + \exp\big(-\lambda \min(x_1, x_2)\big)(1 - H_1(x_1))(1 - H_2(x_2)) \\ &= 1 - (1 - F_1(x_1)) - (1 - F_2(x_2)) + \exp(-\lambda \min(x_1, x_2)) \exp\big(\lambda(x_1 + x_2)\big)(1 - F_1(x_1))(1 - F_2(x_2)).\end{aligned}$$

$$F(x_1,x_2) = F_1(x_1) + F_2(x_2) - 1 + \Big[\big(1 - F_1(x_1)\big)^{-1/\alpha} + \big(1 - F_2(x_2)\big)^{-1/\alpha} - 1\Big]^{-\alpha}.$$

This yields the copula function

$$C(u_1,\ u_2) = u_1 + u_2 - 1 + \Big[\big(1 - u_1\big)^{-1/\alpha} + \big(1 - u_2\big)^{-1/\alpha} - 1\Big]^{-\alpha}. \qquad (1.3)$$

With this function, we can express the bivariate distribution function as $H(x_1,\ x_2) = C\big(F_1(x_1),\ F_2(x_2)\big)$.

Alternatively, we can consider the copula

$$C_*(u_1,\ u_2) = C(1 - u_1,\ 1 - u_2) = \big(u_1^{-1/\alpha} + u_2^{-1/\alpha} - 1\big)^{-\alpha} - 1$$

and express the joint survival function as $\mathrm{Prob}\big(X_1 > x_1,\ X_2 > x_2\big) = C_*\big(S_1(x_1),\ S_2(x_2)\big)$, where $S = 1 - F$. Because our motivating examples in Section 2 concern lifetime (positive) random variables, we often find it intuitively appealing to work with survival in lieu of distribution functions.

Several methods are available for constructing multivariate distributions; see Hougaard (1987) and Hutchinson and Lai (1990) for detailed reviews. Example 1.1 illustrates the so-called "variables-in-common" technique in which a common element serves to induce dependencies among several random variables. This article focuses on the compounding method illustrated in Example 1.2 for two reasons. First, there is a long history of using compound distributions for risk classification in the actuarial science literature, particularly within the credibility framework. Second, Marshall and Olkin (1988) showed that compounding can be used to generate several important families of copulas. Additional discussion of this point appears in Sections 2 and 3.

Examples 1.1 and 1.2 each describe bivariate distributions through probabilistic interpretations of random quantities. It is also useful to explore (in Section 3) a class of functions called "Archimedean copulas," which arise from the mathematical theory of associativity. An important special case of this class, due to Frank (1979), is

$$C(u,\ v) = \frac{1}{\alpha} \ln \left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^{\alpha} - 1}\right). \qquad (1.4)$$

Although Frank's copula does not appear to have a natural probabilistic interpretation, its other desirable properties make it well suited for empirical applications (Nelson 1986 and Genest 1987).

The purpose of this paper is to introduce copulas, their characteristics and properties, and their applicability to specific situations. Section 2 reviews empirical applications of copulas in analyzing survival of multiple lives and competing risks. Both are familiar topics to actuaries. Section 3 discusses properties and characteristics of copulas. In particular, we show (1) how to specify a copula, (2) how the association structure of copulas can be summarized in terms of familiar measures of dependence, and (3) how simulation of multivariate outcomes can be easily accomplished when the distribution is expressed as a copula. Section 4 provides an illustration of fitting a copula to insurance company losses and expenses. Section 5 reviews additional applications of copulas. We conclude in Section 6.

2. EMPIRICAL APPLICATIONS

Copulas are useful for examining the dependence structure of multivariate random vectors. In this section, we describe two biological science subject areas that are related to actuarial science and that have used copulas to understand empirical relationships among multivariate observations.

2.1 Survival of Multiple Lives

In epidemiological and actuarial studies, it is often of interest to examine the joint mortality pattern of groups of more than a single individual. This group

[2] $F(x_1,x_2) = 1 - \mathrm{Prob}(X_1 > x_1) - \mathrm{Prob}(X_2 > x_2) + \mathrm{Prob}(X_1 > x_1,\ X_2 > x_2)$

$$= 1 - \left(1 + \frac{x_1}{\lambda}\right)^{-\alpha} - \left(1 + \frac{x_2}{\lambda}\right)^{-\alpha} + \int \mathrm{Prob}(X_1 > x_1|\gamma)\ \mathrm{Prob}(X_2 > x_2|\gamma)\ \frac{\alpha^{\lambda}}{\Gamma(\lambda)}\ \gamma^{\alpha-1}\ e^{-\lambda\gamma}\ d\gamma$$

$$= 1 - \left(1 + \frac{x_1}{\lambda}\right)^{-\alpha} - \left(1 + \frac{x_2}{\lambda}\right)^{-\alpha} + \int e^{-\gamma x_1}\ e^{-\gamma x_2}\ \frac{\alpha^{\lambda}}{\Gamma(\lambda)}\gamma^{\alpha-1}\ e^{-\lambda\gamma}\ d\gamma$$

$$= 1 - \left(1 + \frac{x_1}{\lambda}\right)^{-\alpha} - \left(1 + \frac{x_2}{\lambda}\right)^{-\alpha} + \left[1 + \frac{x_1 + x_2}{\lambda}\right]^{-\alpha}.$$

could be, for example, a husband and wife, a family with children, or twins (identical or nonidentical). There is strong empirical evidence that supports the dependence of mortality on pairs of individuals. For example, statistical analyses of mortality patterns of married couples are frequently made to test the "broken heart" syndrome. Using a dataset consisting of 4,486 55-year-old widowers, Parkes et al. (1969) showed that there is a 40% increase in mortality among the widowers during the first few months after the death of their wives; see also Ward (1976). Intuitively, pairs of individuals exhibit dependence in mortality because they share common risk factors. These factors may be purely genetic, as in the case of twins, or environmental, as in the case of a married couple.

The first application of copulas in joint-life models arose indirectly through the work of Clayton (1978) in his study of bivariate life tables of fathers and sons. Clayton developed the bivariate distribution function given in Equation (1.3) as the solution of a second-order partial differential equation. Clayton also pointed out the random effects interpretation of the model that was subsequently developed by Oakes (1982). See also Cook and Johnson (1981).

Random effects models are important in biological and epidemiological studies because they provide a method of modeling heterogeneity. A random effects model particularly suited for multivariate survival analysis is the *frailty* model, due to Vaupel, Manton and Stallard (1979) and Hougaard (1984). To describe frailty models, we first introduce some notation. In survival analysis, it is customary to consider the complement of the distribution function, the survival function, and the negative derivative of its logarithmic transform, the hazard function. Thus, for a continuous random survival time T, we define

$$S(t) = \text{Prob}(T > t) = 1 - F(t)$$

and

$$h(t) = -\frac{\partial \ln S(t)}{\partial t} = \frac{f(t)}{S(t)}.$$

Actuaries know the hazard function $h(t)$ as the force of mortality (see, for example, Bowers et al. 1997, Chap. 3).

To understand explanatory variables Z in survival analysis, we can use the Cox (1972) proportional hazards model, which represents the hazard function as

$$h(t, Z) = e^{\beta Z}\, b(t),$$

where $b(t)$ is the so-called "baseline" hazard function and β is a vector of regression parameters. It is proportional in the sense that all the information contained in the explanatory variables is in the multiplicative factor $\gamma = e^{\beta Z}$. Integrating and exponentiating the negative hazard, we can also express Cox's proportional hazard model as

$$S(t|\gamma) = \exp\Big(-\int_0^t h(s, Z)\,ds\Big) = B(t)^{\gamma}.$$

Here,

$$B(t) = \exp\Big(-\int_0^t b(s)ds\Big)$$

is the survival function corresponding to the baseline hazard. Frailty models arise when Z, and hence γ, is unobserved. The factor γ is called a *frailty* because larger values of γ imply a smaller survival function, $S(t|\gamma)$, indicating poorer survival. As demonstrated in Example 1.2, the marginal distribution for a single life T is obtained by taking expectations over the potential values of γ; that is, $S(t) = E_{\gamma} S(t|\gamma)$.

Oakes (1989, 1994) described how frailties can be used to model the dependencies among multiple lives. Other studies, such as Jagger and Sutton (1991), used a Cox regression model with *known* explanatory variables Z to account for the dependencies among multiple lives. Multivariate frailty models are obtained when the investigator does not wish to attribute, or does not have knowledge of, specific characteristics that may induce dependencies. For multivariate frailty models, we assume that "p" lives $T_1, T_2, \ldots, T_p$ are independent given the frailty γ. That is,

$$\begin{aligned}
&\text{Prob}\big(T_1 > t_1, \ldots, T_p > t_p|\gamma\big) \\
&\qquad = \text{Prob}\big(T_1 > t_1|\gamma\big) \cdots \text{Prob}\big(T_p > t_p|\gamma\big) \\
&\qquad = S_1(t_1|\gamma) \cdots S_p(t_p|\gamma) \\
&\qquad = B_1(t_1)^{\gamma} \cdots B_p(t_p)^{\gamma}.
\end{aligned}$$

The joint multivariate survival function is defined as

$$\begin{aligned}
&\text{Prob}\big(T_1 > t_1, \ldots, T_p > t_p\big) \\
&\qquad = E_{\gamma}\Big\{B_1(t_1) \cdots B_p(t_p)\Big\}^{\gamma}. \qquad (2.1)
\end{aligned}$$

Example 2.1 Hougaard's Copula Family

To illustrate, an important frailty model was given by Hougaard (1986), who assumed that the distribution of γ could be modeled as a positive "stable distribution" with Laplace transform $E_{\gamma} e^{-s\gamma} = \exp(-s^{\alpha})$

and parameter α. Recall that the Laplace transform of a positive random variable γ is defined by

$$\tau(s) = E_\gamma \, e^{-s\gamma} = \int e^{-st} \, dG_\gamma(t),$$

where G_γ is the distribution function of γ. This is also the moment generating function evaluated at -s; thus, knowledge of $\tau(s)$ determines the distribution.

With a positive stable distribution for γ, using Equation (2.1) we have

$$\begin{aligned}\text{Prob}\Big(T_1 > t_1, \ldots, T_p > t_p\Big) &= E_\gamma \exp\Big(\gamma \ln \Big\{B_1(t_1) \cdots B_p(t_p)\Big\}\Big) \\ &= \exp\Big(-\Big\{-\ln B_1(t_1) - \cdots - \ln B_p(t_p)\Big\}^\alpha\Big).\end{aligned}$$

Because

$$S_i(t_i) = \exp\Big(-\Big\{-\ln B_i(t_i)\Big\}^\alpha\Big),$$

we can write the joint survival function as

$$\begin{aligned}&\text{Prob}\Big(T_1 > t_1, \ldots, T_p > t_p\Big) \\ &= \exp\Big(-\Big[\Big\{- \ln S_1(t_1)\Big\}^{1/\alpha} + \cdots + \Big\{- \ln S_p(t_p)\Big\}^{1/\alpha}\Big]^\alpha\Big),\end{aligned} \tag{2.2}$$

a copula expression. In particular, for bivariate lifetimes with $p=2$, Hougaard proposed examining Weibull marginals so that $B_i(t)=\exp(-a_i t^{b_i})$ and $S_i(t|\gamma)= \exp(-a_i \gamma t^{b_i})$. This yields the bivariate survivor function

$$\begin{aligned}&\text{Prob}\Big(T_1 > t_1, T_2 > t_2\Big) \\ &\qquad = \exp\Big(-\Big[a_1 t_1^{b_1} + a_2 t_2^{b_2}\Big]^\alpha\Big).\end{aligned} \tag{2.3}$$

This is desirable in the sense that both the conditional and marginal distributions are Weibull.

Equations (1.2) of Example 1.2 and (2.2) of Example 2.1 show that these frailty models can be written as copulas. Marshall and Olkin (1988) showed that these are special cases of a more general result; they demonstrated that all frailty models of the form in Equation (2.1) can be easily written as copulas. Further, the copula form is a special type called an Archimedean copula, which we will introduce in Section 3.

In addition to the Clayton and Oakes studies, other works have investigated the use of copula models in studying behavior of multiple lives. Hougaard et al. (1992) analyzed the joint survival of Danish twins born between 1881 and 1930. They use the frailty model arising from a positive stable distribution as well as Cox's proportional hazard model. Frees et al. (1995) investigated mortality of annuitants in joint- and last-survivor annuity contracts using Frank's copula (Equation 1.3). They found that accounting for dependency in mortality produced approximately a 3% to 5% reduction in annuity values when compared to standard models that assume independence.

2.2 Competing Risks—Multiple Decrement Theory

The subject of competing risks deals with the study of the lifetime distribution of a system subject to several competing causes; this subject is called multiple decrement theory in actuarial science (see, for example, Bowers et al. 1997, Chap. 10 and 11). The problem of competing risks arises in survival analysis, systems reliability, and medical studies as well as in actuarial science. For example, a person dies because of one of several possible causes: cancer, heart disease, accident, and so on. As yet another example, a mechanical device fails because a component fails. For mathematical convenience, the general framework begins with an unobserved multivariate lifetime vector (T_1, T_2, . . . , T_p); each element in the vector denotes the lifetime due to one of p competing causes. The quantities typically observed are $T=\min(T_1, T_2, \ldots, T_p)$ and the cause of failure J. To illustrate, in life insurance, T usually denotes the lifetime of the insured individual and J denotes the cause of death such as cancer or accidental death. Several texts lay the foundation of the theory of competing risks. For example, see Bowers et al. (1997), Cox and Oakes (1984), David and Moeschberger (1978), and Elandt-Johnson and Johnson (1980).

When formulating the competing risk model, it is often assumed that the component lifetimes T_i are statistically independent. With independence, the model is easily tractable and avoids the problem of identifiability encountered in inference. However, many authors, practitioners, and academicians recognize that this assumption is not practical, realistic, or reasonable; see Carriere (1994), Makeham (1874), and Seal (1977).

To account for dependence in competing risk models, one general approach is to apply copulas. In particular, the frailty model seems well suited for handling competing risks. Assuming that causes of death are independent given a frailty γ, we have

$$\begin{aligned}\text{Prob}(T > t|\gamma) &= \text{Prob}\Big(\min(T_1, \ldots, T_p) > t\Big)\\ &= \text{Prob}(T_1 > t|\gamma) \cdots \text{Prob}(T_p > t|\gamma)\\ &= B_1(t)^\gamma \cdots B_p(t)^\gamma.\end{aligned}$$

Thus, similar to Equation (2.1), the overall survival function is

$$\text{Prob}(T > t) = E_\gamma\Big\{B_1(t) \cdots B_p(t)\Big\}^\gamma. \qquad (2.4)$$

Example 2.1 (Continued)

For a positive stable distribution for γ, the survival function is

$$\text{Prob}(T > t) = \exp\Big(-\Big[\Big\{-\ln S_1(t)\Big\}^{1/\alpha} + \cdots + \Big\{-\ln S_p(t)\Big\}^{1/\alpha}\Big]^\alpha\Big),$$

similar to Equation (2.2). For bivariate lifetimes with Weibull marginals, we have

$$\text{Prob}(T > t) = \exp\Big(-\Big[a_1 t^{b_1} + a_2 t^{b_2}\Big]^\alpha\Big).$$

There have been several applications of frailty models for studying competing risk situations. Oakes (1989) discussed the number of cycles of two chemotherapy regimes tolerated by 109 cancer patients. Shih and Louis (1995) analyzed HIV-infected patients by using Clayton's family, positive stable frailties, as well as Frank's copula. Zheng and Klein (1995) considered data from a clinical trial of patients with non-Hodgkin's lymphoma using gamma copula (as in Clayton's family). In a nonbiological context, Hougaard (1987) described how dependent competing risk models using positive stable copulas can be used to assess machine failure.

3. PROPERTIES OF COPULAS

This section discusses several properties and characteristics of copulas, specifically (1) how to generate copulas, (2) how copulas can summarize association between random variables, and (3) how to simulate copula distributions.

3.1 Specifying Copulas: Archimedean and Compounding Approaches

Copulas provide a general structure for modeling multivariate distributions. The two main methods for specifying a family of copulas are the Archimedean approach and the compounding approach, the latter illustrated in Example 1.2.

The Archimedean representation allows us to reduce the study of a multivariate copula to a single univariate function. For simplicity, we first consider bivariate copulas so that $p=2$. Assume that ϕ is a convex, decreasing function with domain (0, 1] and range $[0, \infty)$ such that $\phi(1)=0$. Use ϕ^{-1} for the inverse function of ϕ. Then the function

$$C_\phi(u,v) = \phi^{-1}\,(\phi(u) + \phi(v)) \text{ for } u, v \in (0, 1]$$

is said to be an Archimedean copula. We call ϕ a *generator* of the copula C_ϕ. Genest and McKay (1986a, 1986b) give proofs of several basic properties of C_ϕ, including the fact that it is a distribution function. As seen in Table 1, different choices of generator yield several important families of copulas. A generator uniquely determines (up to a scalar multiple) an Archimedean copula. Thus, this representation helps identify the copula form. This point is further developed in Section 3.2.

Examples 1.2 and 2.1 show that compound distributions can be used to generate copulas of interest.

TABLE 1
ARCHIMEDEAN COPULAS AND THEIR GENERATORS

Family	Generator $\phi(t)$	Dependence Parameter (α) Space	Bivariate Copula $C_\phi(u,v)$
Independence	$-\ln t$	Not applicable	uv
Clayton (1978), Cook-Johnson (1981), Oakes (1982)	$t^{-\alpha} - 1$	$\alpha > 1$	$\Big(u^{-\alpha} + v^{-\alpha} - 1\Big)^{-1/\alpha}$
Gumbel (1960), Hougaard (1986)	$\Big(-\ln t\Big)^\alpha$	$\alpha \geq 1$	$\exp\Big\{-\Big[\Big(-\ln u\Big)^\alpha + \Big(-\ln v\Big)^\alpha\Big]^{1/\alpha}\Big\}$
Frank (1979)	$\ln \dfrac{e^{\alpha t} - 1}{e^\alpha - 1}$	$-\infty < \alpha < \infty$	$\dfrac{1}{\alpha}\ln\Big(1 + \dfrac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^\alpha - 1}\Big)$

These examples are special cases of a general method for constructing copulas due to Marshall and Olkin (1988). To describe this method, suppose that X_i is a random variable whose conditional, given a positive latent variable γ_i, distribution function is specified by $H_i(x|\gamma_i) = H_i(x)^{\gamma_i}$, where $H_i(\cdot)$ is some baseline distribution function, for $i=1, \ldots, p$. Marshall and Olkin considered multivariate distribution functions of the form

$$F(x_1, x_2, \ldots, x_p) = EK\Big(H_1(x_1)^{\gamma_1}, \ldots, H_p(x_p)^{\gamma_p}\Big).$$

Here, K is a distribution function with uniform marginals and the expectation is over $\gamma_1, \gamma_2, \ldots, \gamma_p$. As a special case, take all latent variables equal to one another so that $\gamma_1 = \gamma_2 = \ldots = \gamma_p = \gamma$ and use the distribution function corresponding to independent marginals. Marshall and Olkin (1988) showed that

$$\begin{aligned} F\ x_1,\ x_2, \ldots, x_p) &\\ = E_\gamma\Big(H_1(x_1)^\gamma \cdots H_p(x_p)^\gamma\Big) &\\ = \tau\Big(\tau^{-1}\Big\{F_1(x_1)\Big\} + \cdots + \tau^{-1}\Big\{F_p(x_p)\Big\}\Big) & \quad (3.2) \end{aligned}$$

where F_i is the i-th marginal distribution of F and $\tau(\cdot)$ is the Laplace transform of γ, defined by $\tau(s) = E_\gamma e^{-s\gamma}$.

Laplace transforms have well-defined inverses. Thus, from Equation (3.2), we see that the inverse function τ^{-1} serves as the generator for an Archimedean copula. In this sense, Equation (3.2) provides a probabilistic interpretation of generators. To illustrate, Table 2 provides the inverse Laplace transform for the generators listed in Table 1. Here, we see how well-known distributions can be used to generate compound distributions. Because generators are defined uniquely only up to scalar multiple, any positive constant in the family of Laplace transforms determines the same class of generators. (Indeed, this methodology suggests new copula families.) Thus, the inverse of a Laplace transform represents an important type of generator for Archimedean copulas.

To summarize, assume that $X_1, X_2, \ldots, X_p$ are conditionally, given γ, independent with distribution functions $H_i(x)^\gamma$. Then, the multivariate distribution is given by the copula form with the generator being the inverse of the Laplace transform of the latent variable γ. Because of the form of the conditional distribution, we follow Joe (1997) and call this a *mixture of powers* distribution. We remark that

$$\tau\big[-\ln H_i(x)\big] = E_\gamma \exp\big\{-\big[-\ln H_i(x)\big]\gamma\big\} = F_i(x)$$

so that

$$H_i(x) = \exp\big\{-\tau^{-1}\big[F_i(x)\big]\big\}.$$

This provides a way of specifying the baseline function given the marginal distribution and the distribution of the latent variable.

For the applications involving lifetimes in Section 2, we found it natural to discuss distributions in terms of survival functions. Marshall and Olkin (1988) pointed out that the construction as in Equation (3.2) could also be used for survivor functions. That is, with the frailty model $\text{Prob}\big(T_i > t|\gamma\big) = B_i(t)^\gamma$ and conditional independence of $T_1, T_2, \ldots, T_p$, from Equation (2.1), we have

$$\begin{aligned} \text{Prob}(T_1 > t_1, \ldots, T_p > t_p) &\\ = E_\gamma\Big\{B_1(t_1) \cdots B_p(t_p)\Big\}^\gamma &\\ = \tau\Big(\tau^{-1}\Big\{S_1(x_1)\Big\} &\\ + \cdots + \tau^{-1}\Big\{S_p(x_p)\Big\}\Big) & \quad (3.3) \end{aligned}$$

TABLE 2
ARCHIMEDEAN GENERATORS AND THEIR INVERSES

Family	Generator $\phi(t)$	Inverse Generator (Laplace Transform) $\tau(s) = \phi^{-1}(s)$	Laplace Transform Distribution
Independence	$-\ln t$	$\exp(-s)$	Degenerate
Clayton (1978), Cook-Johnson (1981), Oakes (1982)	$t^{-\alpha} - 1$	$(1 + s)^{-1/\alpha}$	Gamma
Gumbel (1960), Hougaard (1986)	$(-\ln t)^\alpha$	$\exp(-s^{1/\alpha})$	Positive stable
Frank (1979)	$\ln \dfrac{e^{\alpha t} - 1}{e^\alpha - 1}$	$\alpha^{-1} \ln\big[1 + e^s(e^\alpha - 1)\big]$	Logarithmic series distribution on the positive integers

As before,

$$\tau\left[-\ln B_i(t)\right] = S_i(t) = 1 - F_i(t)$$

and

$$B_i(t) = \exp\left\{-\tau^{-1}\left[S_i(t)\right]\right\}.$$

If the mixing distribution remains the same, the Laplace transform and hence the generator remain the same. However, the two constructions yield different multivariate distributions because

$$\text{Prob}\left(T \le t|\gamma\right) = 1 - B(t)^{\gamma} \neq \left(1 - F(t)\right)^{\gamma}.$$

We follow Marshall and Olkin and first present the copula construction in Equation (3.2) by using distribution functions because it is useful for all random variables. However, for positive lifetime random variables, the concept of frailty models is intuitively appealing; thus, using survival functions in the construction is preferred for these applications. We finally remark, unlike the gamma and positive stable families, that Frank's copula $C_\phi(u,v)$ is symmetric about the point (1/2, 1/2) [for example, $C_\phi(u,v) = C_\phi(1/2 - u, 1/2 - v)$]. Thus, it is invariant to the choice of F or $S = 1 - F$ in the construction (Genest 1987).

3.2 Measures of Association

Recall the copula representation of a distribution function in Equation (1.1),

$$F(x_1, x_2, \ldots, x_p) = C\left(F_1(x_1), F_2(x_2), \ldots, F_p(x_p)\right).$$

With this expression, we see that F is a function of its marginals and the copula. As pointed out by Genest and Rivest (1993), this suggests that a natural way of specifying the distribution function is to examine the copula and marginals separately. Moreover, the case of independence is a special form of the copula $C(u_1, u_2, \ldots, u_p) = u_1 \cdot u_2 \cdots u_p$ (regardless of the marginals), and this suggests that we examine the copula function to understand the association among random variables. Because we are concerned with correlation measures, we restrict our consideration to $p = 2$.

Schweizer and Wolff (1981) established that the copula accounts for all the dependence between two random variables, X_1 and X_2, in the following sense. Consider g_1 and g_2, strictly increasing (but otherwise arbitrary) functions over the range of X_1 and X_2. Then, Schweizer and Wolff showed that the transformed variables $g_1(X_1)$ and $g_2(X_2)$ have the same copula as X_1 and X_2. Thus, the manner in which X_1 and X_2 "move together" is captured by the copula, regardless of the scale in which each variable is measured.

Schweizer and Wolff also showed that two standard nonparametric correlation measures could be expressed solely in terms of the copula function. These are Spearman's correlation coefficient, defined by

$$\begin{aligned}\rho(X_1, X_2) &= 12E\left\{\left(F_1(x_1) - 1/2\right)\left(F_2(x_2) - 1/2\right)\right\} \\ &= 12 \iint \left\{C(u,v) - uv\right\} du\,dv\end{aligned}$$

and Kendall's correlation coefficient, defined by

$$\begin{aligned}\tau(X_1, X_2) &= \text{Prob}\left\{(X_1 - X_1^*)(X_2 - X_2^*) > 0\right\} \\ &\quad - \text{Prob}\left\{(X_1 - X_1^*)(X_2 - X_2^*) < 0\right\} \\ &= 4 \iint C(u,v)\,dC(u,v) - 1.\end{aligned}$$

For these expressions, we assume that X_1 and X_2 have a jointly continuous distribution function. Further, the definition of Kendall's τ uses an independent copy of (X_1, X_2), (X_1^*, X_2^*) to define the measure of "concordance." See Section 5 for more details. Also, the widely used Pearson correlation coefficient, $\text{Cov}(X_1, X_2)/(\text{Var}X_1\, \text{Var}X_2)^{1/2}$, depends not only on the copula but also on the marginal distributions. Thus, this measure is affected by (nonlinear) changes of scale.

Table 3 illustrates the calculation of these correlation measures. The correlations from Frank's copula rely on the so-called "Debye" functions, defined as

$$D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t - 1}dt,$$

for $k = 1,2$. To evaluate negative arguments of the Debye function D_k, basic calculus shows that

$$D_k(-x) = D_k(x) + \frac{kx}{k+1}.$$

An important point of this table is that there is a one-to-one relationship between each correlation measure and the association parameter α. Further, Table 3 allows us to see a drawback of the Clayton/Cook-Johnson/Oakes and Gumbel-Hougaard copula families. Because of the limited dependence parameter space as shown in Table 1, these families permit only nonnegative correlations, a consequence of the latent variable model. However, Frank's family permits negative as well as positive dependence.

Correlation measures summarize information in the copula concerning the dependence, or association, between random variables. Following a procedure due to Genest and Rivest (1993), we can also

TABLE 3
Archimedean Copulas and Their Measures of Dependence

Family	Bivariate Copula $C_\phi(u,v)$	Kendall's τ	Spearman's ρ
Independence	$u\,v$	0	0
Clayton (1978), Cook-Johnson (1981), Oakes (1982)	$\left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}$	$\frac{\alpha}{\alpha + 2}$	Complicated form
Gumbel (1960), Hougaard (1986)	$\exp\left\{-\left[(-\ln u)^{\alpha} + (-\ln v)^{\alpha}\right]^{1/\alpha}\right\}$	$1 - \alpha^{-1}$	No closed form
Frank (1979)	$\frac{1}{\alpha}\ln\left(1 + \frac{(e^{\alpha u} - 1)(e^{\alpha v} - 1)}{e^{\alpha} - 1}\right)$	$1 - \frac{4}{\alpha}\{D_1(-\alpha) - 1\}$	$1 - \frac{12}{\alpha}\{D_2(-\alpha) - D_1(-\alpha)\}$

use the dependence measure to specify a copula form in empirical applications, as follows.

Genest and Rivest's procedure for identifying a copula begins by assuming that we have available a random sample of bivariate observations, $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$. Assume that the distribution function F has associated Archimedean copula C_ϕ; we wish to identify the form of ϕ. We work with an intermediate (unobserved) random variable $Z_i = F(X_{1i}, X_{2i})$ that has distribution function $K(z) = \text{Prob}(Z_i \le z)$. Genest and Rivest showed that this distribution function is related to the generator of an Archimedean copula through the expression

$$K(z) = z - \frac{\phi(z)}{\phi'(z)}.$$

To identify ϕ, we:

1. Estimate Kendall's correlation coefficient using the usual (nonparametric or distribution-free) estimate
$$\tau_n = \binom{n}{2}^{-1} \sum_{i<j} \text{sign}\left[(X_{1i} - X_{1j})(X_{2i} - X_{2j})\right].$$
2. Construct a nonparametric estimate of K, as follows:
   a. First, define the pseudo-observations Z_i = {number of (X_{1j}, X_{2j}) such that $X_{1j} < X_{1i}$ and $X_{2j} < X_{2i}$} / $(n - 1)$ for $i = 1, \ldots, n$.
   b. Second, construct the estimate of K as $K_n(z)$ = proportion of Z_i's $\le z$.
3. Now construct a parametric estimate of K using the relationship
$$K_\phi(z) = z - \frac{\phi(z)}{\phi'(z)}.$$
For example, refer to Table 1 for various choices of ϕ and use the estimate τ_n to calculate an estimate of α, say α_n. Use α_n to estimate $\phi(x)$, say $\phi_n(x)$. Finally, use $\phi_n(x)$ to estimate $K_\phi(z)$, say $K_{\phi_n}(z)$.

Repeat step 3 for several choices of ϕ. Then, compare each parametric estimate to the nonparametric estimate constructed in Step 2. Select the choice of ϕ so that the parametric estimate $K_{\phi_n}(z)$ most closely resembles the nonparametric estimate $K_n(z)$. Measuring "closeness" can be done by minimizing a distance such as $\int [K_{\phi_n}(z) - K_n(z)]^2 \, dK_n(z)$ or graphically. Graphical representations include (1) plots of $K_n(z)$ and $K_{\phi_n}(z)$ versus z, and (2) the corresponding quantile plots. See Section 4 for an example.

3.3 Simulation

Actuaries routinely deal with complex nonlinear functions, such as present values, of random variables. Simulation is a widely used tool for summarizing the distribution of stochastic outcomes and for communicating the results of complex models. The copula construction allows us to simulate outcomes from a multivariate distribution easily.

The two primary simulation strategies are the Archimedean and compounding methods for constructing copulas; each has relative advantages and disadvantages.

We begin with the Archimedean construction. Our goal is to construct an algorithm to generate $X_1, X_2, \ldots, X_p$ having known distribution function $F(x_1, x_2, \ldots, x_p) = C(F_1(x_1), F_2(x_2), \ldots, F_p(x_p))$, where the copula function is

$$C(u_1, u_2, \ldots, u_p) = \phi^{-1}\left(\phi(u_1) + \cdots + \phi(u_p)\right).$$

For this construction, Genest and Rivest (1986b) and Genest (1987) introduced the idea of simulating the full distribution of $(X_1, X_2, \ldots, X_p)$ by recursively simulating the conditional distribution of X_k given $X_1, \ldots, X_{k-1}$, for $k=2, \ldots, p$. This idea, subsequently developed by Lee (1993), is as follows. For simplicity, we assume that the joint probability density function of $X_1, X_2, \ldots, X_p$ exists. Using the copula construction, the joint probability density function of $X_1, \ldots, X_k$ is

$$f_k(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \ldots \partial x_k} \phi^{-1}$$

$$\left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_k(x_k)\right]\right\}$$

$$= \phi^{-1(k)} \left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_k(x_k)\right]\right\}$$

$$\prod_{j=1}^{k} \phi^{(1)} \left[F_j(x_j)\right] F_j^{(1)} (x_j).$$

Here, the superscript notation (j) means the j-th mixed partial derivative. Thus, the conditional density of X_k given $X_1, \ldots, X_{k-1}$ is

$$f_k(x_k|x_1, \ldots, x_{k-1}) = \frac{f_k(x_1, \ldots, x_k)}{f_{k-1}(x_1, \ldots, x_{k-1})}$$

$$= \phi^{(1)} \left[F_k(x_k)\right] F^{(1)} (x_k)$$

$$\frac{\phi^{-1(k-1)} \left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_k(x_k)\right]\right\}}{\phi^{-1(k-1)} \left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_{k-1}(x_{k-1})\right]\right\}}.$$

Further, the conditional distribution function of X_k given $X_1, \ldots, X_{k-1}$ is

$$F_k\left(x_k|x_1, \ldots, x_{k-1}\right)$$

$$= \int_{-\infty}^{x_k} f_k \left(x|x_1, \ldots, x_{k-1}\right) dx$$

$$= \frac{\phi^{-1(k-1)} \left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_k(x_k)\right]\right\}}{\phi^{-1(k-1)} \left\{\phi\left[F_1(x_1)\right] + \cdots + \phi\left[F_{k-1}(x_{k-1})\right]\right\}}$$

$$= \frac{\phi^{-1(k-1)} \left\{\mathrm{c}_{k-1} + \phi\left[F_k(x_k)\right]\right\}}{\phi^{-1(k-1)} \left(\mathrm{c}_{k-1}\right)},$$

where $c_k=\phi[F_1(x_1)]+\cdots+\phi[F_k(x_k)]$. With this distribution function, we can now use the usual procedure of solving for the inverse distribution function and evaluating this at a uniform random number; that is, use $F^{-1}(U)=X_k$.

To summarize, the algorithm is (Lee 1993):

Algorithm 3.1 Generating Multivariate Outcomes from an Archimedean Copula

1. Generate $U_1, U_2, \ldots, U_p$ independent uniform (0,1) random numbers.
2. Set $X_1=F_1^{-1}(U_1)$ and $c_0 = 0$.
3. For $k=2, \ldots, p$, recursively calculate X_k as the solution of

$$U_k = F_k\left(X_k|x_1, \ldots, x_{k-1}\right)$$

$$= \frac{\phi^{-1(k-1)} \left\{c_{k-1} + \phi\left[F_k(x_k)\right]\right\}}{\phi^{-1(k-1)} \left(c_{k-1}\right)}. \quad (3.4)$$

This algorithm was initially introduced in the context of Frank's copula for $p=2$. Here, pleasant calculations show that the algorithm reduces to:

Algorithm 3.2 Generating Bivariate Outcomes from Frank's Copula

1. Generate U_1, U_2 independent uniform (0,1) random numbers.
2. Set $X_1=F_1^{-1}(U_1)$.
3. Calculate X_2 as the solution of

$$U_2 = e^{-\alpha U_1} \left(\frac{e^{-\alpha} - e^{-\alpha F_2(X_2)}}{e^{-\alpha F_2(X_2)} - 1} + 1\right)^{-1}.$$

That is, calculate $X_2=F_2^{-1}(U_{*2})$ where

$$U_{*2} = \frac{U_2 e^{-\alpha} - e^{-\alpha U_1} \left(1 - U_{*2}\right)}{U_2 + e^{-\alpha U_1} \left(1 - U_{*2}\right)}.$$

Genest (1987) gave this algorithm.

The algorithm can also be readily used to simulate distributions from Clayton's family. From Table 2, we have $\phi^{-1}(s)=(1+s)^{-1/\alpha}$ so that

$$\phi^{-1(1)}(s)=-\alpha^{-1}(1+s)^{-(1/\alpha)-1}.$$

This expression with $p=2$ and Equation (3.4) show that

$$U_2 = \frac{-\alpha^{-1} \left(1 + F_2(X_2)^{-1/\alpha} - 1 + U_1^{-1/\alpha} - 1\right)^{-(1/\alpha)-1}}{-\alpha^{-1} \left(1 + U_1^{-1/\alpha} - 1\right)^{1(1/\alpha)-1}}.$$

That is, calculate $X_2=F_2^{-1}(U_{*2})$ where

$$U_{*2} = \left(1 + U_1^{-1/\alpha} \left(U_{*2}^{-\alpha/(\alpha+1)} - 1\right)\right)^{-\alpha}.$$

For the Gumbel-Hougaard copula, determining X_2 in Equation (3.4) requires an iterative solution. Although straightforward, this is computationally expensive because many applications require large numbers of simulated values. This drawback leads us to introducing an alternative algorithm suggested by Marshall and Olkin (1988) for compound constructions of copulas.

To generate $X_1, X_2, \ldots, X_p$ having a mixture of powers distribution specified in Equation (3.2), the algorithm is:

Algorithm 3.3 Generating Multivariate Outcomes from a Compound Copula

1. Generate a (latent) random variable γ having Laplace transform τ.
2. Independently of step 1, generate $U_1, U_2, \ldots, U_p$ independent uniform (0, 1) random numbers.
3. For $k=1, \ldots, p$, calculate $X_k=F_k^{-1}(U_{*k})$ where

$$U_{*k} = \tau\left(-\gamma^{-1} \ln U_k\right). \tag{3.5}$$

Recall that the marginal distribution function can be calculated from the baseline distribution function using $F_k(x)=\tau(-\ln H_k(x))$. To illustrate for the Gumbel-Hougaard copula, from Equation (3.5), we have

$$U_{*k} = \exp\left[-\left(-\gamma^{-1} \ln U_k\right)^{-1/\alpha}\right].$$

The algorithm is straightforward for most copulas of interest that are generated by the compounding method. Like the conditional distribution approach, it can be easily implemented for more than two dimensions ($p>2$). It is computationally more straightforward than the conditional distribution approach. A disadvantage is that it requires the generation of an additional variable, γ. For bivariate applications, this means generating 50% more uniform random variates; this can be expensive in applications.

4. Insurance Company Loss and Expense Application

This section illustrates methods of fitting copulas to insurance company indemnity claims. The data comprise 1,500 general liability claims randomly chosen from late settlement lags and were provided by Insurance Services Office, Inc. Each claim consists of an indemnity payment (the loss, X_1) and an allocated loss adjustment expense (ALAE, X_2). Here, ALAE are types of insurance company expenses that are specifically attributable to the settlement of individual claims such as lawyers' fees and claims investigation expenses; see, for example, Hogg and Klugman (1984). Our objective is to describe the joint distribution of losses and expenses.

Estimation of the joint distribution of losses and expenses is complicated by the presence of censoring, a common feature of loss data (Hogg and Klugman 1984). Specifically, in addition to loss and expense information, for each claim we have a record of the policy limit, the maximal claim amount. With the presence of the policy limit, the loss variable is censored because the amount of claim cannot exceed the stated policy limit. For some claims, the policy limit was unknown, and for these policies, we assumed there was no policy limit.

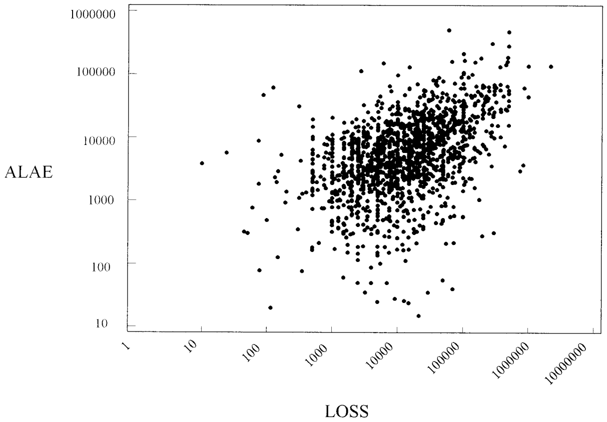
Table 4 summarizes the data. Here, only 34 of 1,500 policies have claims that equaled the policy limit and thus are considered censored. However, the censored losses cannot be ignored; for example, the mean loss of censored claims is much higher than the corresponding mean for uncensored claims. Table 4 also shows that our sample comprises only claims with positive losses and expenses. Separate models would be required for claims with positive losses but no expenses or for claims with zero losses but positive expenses.

Figure 1 is a scatterplot of loss versus ALAE. The corresponding correlation coefficient is 0.41. This

TABLE 4
SUMMARY STATISTICS OF LOSSES AND EXPENSES

	ALAE	Loss	Policy Limit	Loss (Uncensored)	Loss (Censored)
Number	1,500	1,500	1,352	1,466	34
Mean	12,588	41,208	559,098	37,110	217,491
Median	5,471	12,000	500,000	11,048	100,000
Standard Deviation	28,146	102,748	418,649	92,513	258,205
Minimum	15	10	5,000	10	5,000
Maximum	501,863	2,173,595	7,500,000	2,173,595	1,000,000
25th Percentile	2,333	4,000	300,000	3,750	50,000
75th Percentile	12,577	35,000	1,000,000	32,000	300,000

FIGURE 1
PLOT OF ALAE VERSUS LOSS.
BOTH VARIABLES ARE ON A LOGARITHMIC SCALE.
THIRTY-FOUR LOSS OBSERVATIONS ARE CENSORED.
THIS PLOT DEMONSTRATES A STRONG RELATIONSHIP BETWEEN ALAE AND LOSS.



statistic, together with the plot, suggests a strong relationship between losses and expenses.

4.1 Fitting Marginal Distributions

The initial step in our model fitting is to determine the appropriate marginals. Determining appropriate parametric distributions for univariate data is well described in Hogg and Klugman (1984). For these data, this step was examined in detail earlier by Klugman and Parsa (1995). Thus, for simplicity, we present here only the fit of the univariate marginals using a Pareto distribution. With parameters λ and θ, the distribution function is

$$F(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\theta}.$$

The quality of the fit of the marginal distributions can be examined with a graphical comparison of the fitted distribution function against their empirical versions, as displayed in Figures 2 and 3. Because of censoring, we used the Kaplan-Meier empirical distribution function for the loss variable.

4.2 Fitting a Copula to the Bivariate Data

To fit the copula, we first identify the form of the copula in Section 4.2.1 and then estimate it using maximum likelihood in Section 4.2.2.

4.2.1 Identifying a Copula

We use the procedure developed by Genest and Rivest (1993) for identifying an appropriate copula, outlined in Section 3.2. According to the procedure, we examine the degree of closeness of the parametric and nonparametric versions of the distribution function $K(z)$. This procedure is based on estimates of $K(z)$, the distribution function of pseudo-observations $Z=F(X_1, X_2)$. The idea behind the procedure is to compare a

FIGURE 2
FITTED DISTRIBUTION FUNCTIONS OF ALAE.
THE DOTTED CURVE IS THE EMPIRICAL DISTRIBUTION FUNCTION.
THE SMOOTH CURVE IS A FIT USING THE PARETO DISTRIBUTION.

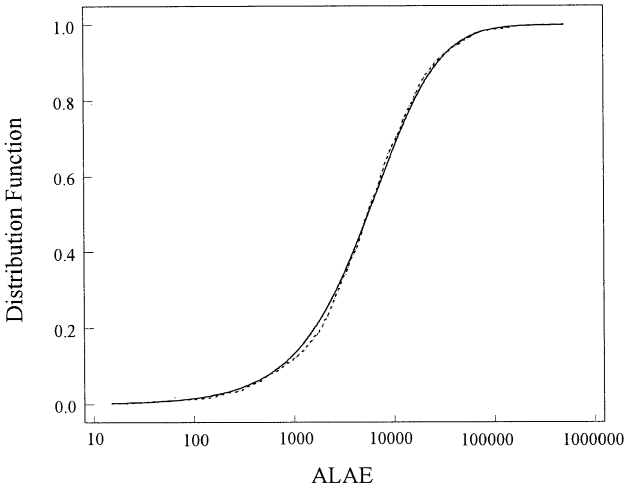
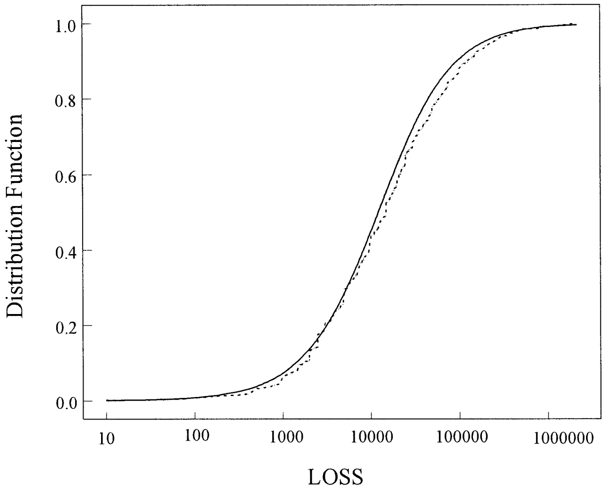


FIGURE 3
FITTED DISTRIBUTION FUNCTIONS OF LOSS.
THE DOTTED CURVE IS A KAPLAN-MEIER EMPIRICAL DISTRIBUTION FUNCTION. THE SMOOTH CURVE IS A FIT USING THE PARETO DISTRIBUTION.



nonparametric estimate of $K(z)$ to those based on a specific form of the copula. To compare the two estimates of $K(z)$, we examine quantiles from each estimated distribution. Scatterplots of the two sets of quantiles are widely known in statistics as quantile-quantile, or *q-q*, plots. For identification purposes, we ignore the mild censoring in the loss variable although we do accommodate it in the more formal maximum likelihood fitting in Section 4.2.2.

We now discuss identification for three widely used copulas, namely, the Gumbel-Hougaard, Frank, and Cook-Johnson copulas. The form of the copula for each of these families appears in Table 3. The comparison of the resulting *q-q* plots is displayed in Figure 4. Because of the close agreement between nonparametric and parametric quantiles, the procedure suggests the use of the Gumbel-Hougaard copula. The quantiles based on Frank's copula are also close to the nonparametric quantiles, although there is greater disparity at the higher quantiles, corresponding to high losses and expenses. We therefore identify both Frank's and Gumbel-Hougaard's as copulas that we fit more formally in Section 4.2.2.

4.2.2 Fitting a Copula Using Maximum Likelihood

Recall that our data consist of losses (X_1) and expenses (X_2) and that we also have available an indicator for censoring (δ), so that $\delta = 1$ indicates the claim is censored. Parameters were estimated using maximum likelihood procedures that were programmed using the SAS procedure IML. In the development of the likelihood equation, we use the following partial derivatives:

$$F_1(x_1, x_2) = \frac{\partial F(x_1, x_2)}{\partial x_1},$$

$$F_2(x_1, x_2) = \frac{\partial F(x_1, x_2)}{\partial x_2},$$

and

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1\, \partial x_2}.$$

Similarly, the first partial derivatives for the copula will be denoted by C_1 and C_2; the second mixed partial derivative by C_{12}.

Using a one-parameter copula and Pareto marginals, we have a total of five parameters to estimate: two each for the marginals and one for the dependence parameter. To develop the likelihood, we distinguish between the censored and uncensored cases. If the loss variable is not censored, then $\delta=0$ and the contribution to the likelihood function is

$$f(x_1, x_2) = f_1(x_1)\, f_2(x_2) C_{12}\left[F_1(x_1), F_2(x_2)\right]. \quad (4.1)$$

If the loss variable is censored, then $\delta=1$ and the joint probability is given by

$$\text{Prob}\left(X_1 \geq x_1, X_2 \leq x_2\right) = F_2(x_2) - F(x_1, x_2).$$

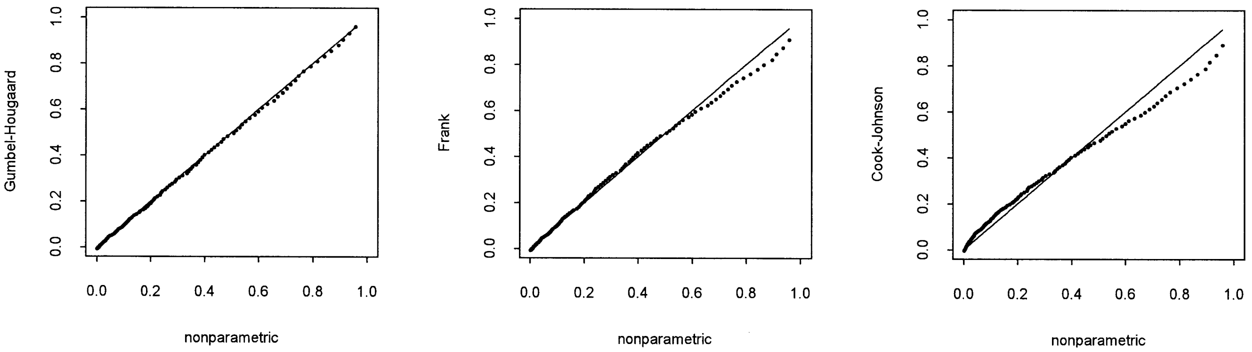
Thus, the contribution to the likelihood when the observation is censored is

$$\begin{aligned} &f_2(x_2) - F_2(x_1, x_2) \\ &\quad = f_2(x_2)\left\{1 - C_2\left[F_1(x_1), F_2(x_2)\right]\right\}. \quad (4.2)\end{aligned}$$

Combining Equations (4.1) and (4.2), for a single observation, the logarithm of the likelihood function is

FIGURE 4

Quantile-quantile (Q-Q) plots, corresponding to the parametric and nonparametric distribution estimates of the pseudo-observations defined in Section 3.2. The dotted lines correspond to the quantiles of nonparametric and parametric estimates of the Archimedean generator ϕ. The smoothed lines correspond to the case where the quantiles are equal.



$$\log L(x_1,\ x_2,\ \delta) = (1 - \delta)\log f(x_1,\ x_2) + \delta\Big\{\log f_2(x_2) + \log\Big[1 - C_2\Big(F_1(x_1),\ F_2(x_2)\Big)\Big]\Big\}. \qquad (4.3)$$

The parameter estimates are then determined by maximizing the likelihood for the entire dataset:

$$\sum_{i=1}^{n} \log L(x_{1_i},\ x_{2_i},\ \delta).$$

Results of the maximum likelihood estimation fitting the Gumbel-Hougaard copula, whose partial derivatives are derived in Appendix A, are summarized in Table 5. Here, we see that the parameter estimates of the marginal distributions are largely unchanged when we compare the univariate to the bivariate estimation. Standard errors are smaller in the bivariate fit, indicating greater precision of the parameter estimates. The estimate of the dependence parameter is significantly different from 1; the estimate of α is approximately 13 standard errors above 1. This provides strong statistical evidence that losses and expenses are not independent. Using Table 3 for the Gumbel-Hougaard copula, we can convert the dependence parameter into a more familiar measure of association. Thus, the parameter value of $\hat{\alpha} = 1.453$ corresponds to an approximate Spearman's correlation measure of 31%. A 95% confidence interval for α is therefore given by $\hat{\alpha} \pm 1.96{*}se(\hat{\alpha}) = 1.453 \pm 1.96(0.034) = (1.386, 1.520)$. This corresponds to a 95% confidence interval of (28%, 34%) for the Spearman's correlation.

Results of the maximum likelihood estimation fitting the Frank copula are summarized in Table 6. Here, the behavior of parameter estimates and standard errors is similar to the fit using the Gumbel-Hougaard copula. For our parameterization of Frank's copula, the case of independence corresponds to $\alpha = 0$. In our case, the parameter estimate of $\hat{\alpha} = -3.158$ corresponds to an approximate Spearman's correlation measure of 32%, which is close to the estimate using the Gumbel-Hougaard copula.

It is difficult to compare the fit of the two copulas directly because they are non-nested models. However, we did compute Akaike's Information Criteria (AIC) for each model, given by AIC=[−2 ln (maximized likelihood)+2(5)]/1500. The results are 15.02 and 15.06 for the Gumbel-Hougaard and Frank copula models, respectively. The smaller AIC value for the Gumbel-Hougaard model indicates that this model is preferred.

4.3 Uses of the Bivariate Fit

This section describes two applications using the estimated bivariate distribution of losses and expenses.

TABLE 5

BIVARIATE DATA PARAMETER ESTIMATES USING GUMBEL-HOUGAARD'S COPULA WITH PARETO MARGINALS

		Bivariate Distribution		Univariate Distribution	
	Parameter	Estimate	Standard Error	Estimate	Standard Error
Loss (X_1)	λ_1	14,036	1,298	14,453	1,397
	θ_1	1.122	0.062	1.135	0.066
ALAE (X_2)	λ_2	14,219	1,426	15,133	1,633
	θ_2	2.118	0.153	2.223	0.175
Dependence	α	1.453	0.034	Not applicable	Not applicable

TABLE 6

BIVARIATE DATA PARAMETER ESTIMATES USING FRANK'S COPULA WITH PARETO MARGINALS

		Bivariate Distribution		Univariate Distribution	
	Parameter	Estimate	Standard Error	Estimate	Standard Error
Loss (X_1)	λ_1	14,558	1,390	14,453	1,397
	θ_1	1.115	0.065	1.135	0.066
ALAE (X_2)	λ_2	16,678	1,824	15,133	1,633
	θ_2	2.309	0.187	2.223	0.175
Dependence	α	−3.158	0.174	Not applicable	Not applicable

4.3.1 Calculating Reinsurance Premiums

After having identified the joint distribution of (X_1, X_2), we can examine the distribution of any known function of X_1 and X_2, say, $g(X_1, X_2)$. To illustrate, consider a reinsurer's expected payment on a policy with limit L and insurer's retention R. Then, assuming a pro-rata sharing of expenses, we have

$$g(X_1, X_2) = \begin{cases} 0 & \text{if } X_1 < R \\ X_1 - R + \dfrac{X_1 - R}{X_1} X_2 & \text{if } R \leq X_1 < L. \\ L - R + \dfrac{L - R}{L} X_2 & \text{if } X_1 \geq L \end{cases}$$

The expected payment $Eg(X_1, X_2)$ could be calculated using numerical integration when the joint density of losses and expenses is available. However, simulation is a simpler, numerical evaluation tool. The procedure for simulation is described in Section 3.3.

The idea with simulation is to generate a sequence of bivariate data (x_{1_i}, x_{2_i}) from the bivariate distribution model. The procedure for the Gumbel-Hougaard copula is summarized in Algorithm 3.3, using Equation (3.6) for Equation (3.5). In the procedure, the inverse of the marginal distribution functions is needed. In the Pareto case, it is not difficult to verify that $F^{-1}(x) = \lambda[(1-x)^{-1/\theta} - 1]$.

The simulation steps outlined in Algorithm 3.3, using the Gumbel-Hougaard copula, are repeated a large number of times. Let $nsim$ be the number of simulations performed so that we generate the sequence of sample (x_{1i}, x_{2i}), $i=1, \ldots, nsim$. Thus, the estimated value for the reinsurer's expected payment is given by

$$\hat{g}^*(L, R) = \frac{1}{nsim} \sum_{i=1}^{nsim} g(x_{1i}, x_{2i}),$$

with standard error

$$se\big(\hat{g}^*(L, R)\big) = \sqrt{\frac{\dfrac{1}{nsim} \displaystyle\sum_{i=1}^{nsim} g(x_{1i}, x_{2i})^2 - \hat{g}^*(L, R)^2}{nsim}}.$$

We performed a simulation study of size $nsim =$ 100,000; the results are summarized in Table 7.

The results in Table 7 provide the adjusted premiums the reinsurer would have assessed to cover costs of losses and expenses according to various policy limits and ratios of insurer's retention to policy limit. On a crude basis, these results appear to make sense. For example, from the summary statistics in Table 4, the average policy limit is 559,098 with an average of losses plus expenses of 53,796. Without any reinsurance, our results indicate an adjusted premium of 49,367, with a standard error of 733, for a policy limit of 500,000. Furthermore, the results are intuitively appealing because as expected we observe (1) higher premium for larger policy limits and (2) lower premium when the ratio R/L is higher, that is, insurer retains larger amount of losses.

Because it is common practice to assume independence, we provide Table 8, which gives ratios of dependence to independence reinsurance premiums. The dependence assumption is based on the Gumbel-Hougaard estimation results, while the independence assumption is based on the estimation results when the joint distribution is assumed to be the product of the marginals. In previous sections, we argue that the estimation results using dependence are statistically significant. Table 8 shows that substantial mispricing could result if the unrealistic assumption of independence between losses and expenses is made. A ratio below 1.0 from the table suggests an overvalued reinsurance premium; a ratio above 1.0 suggests undervalued premiums. According to the

TABLE 7
SIMULATION-BASED REINSURANCE PREMIUMS (SIMULATION STANDARD ERRORS ARE IN PARENTHESIS)

	Ratio of Insurer's Retention to Policy Limit (*R*/*L*)				
Policy Limit (*L*)	0.00	0.25	0.50	0.75	0.95
10,000	15,636 (640)	11,232 (480)	7,220 (320)	3,498 (160)	684 (32)
100,000	34,264 (655)	17,965 (493)	10,003 (328)	4,425 (164)	819 (33)
500,000	49,367 (733)	17,457 (544)	9,234 (359)	4,007 (179)	739 (36)
1,000,000	55,683 (818)	16,762 (597)	8,740 (390)	3,716 (193)	672 (38)

TABLE 8
SIMULATION-BASED RATIOS OF DEPENDENCE TO INDEPENDENCE REINSURANCE PREMIUMS

Policy Limit (L)	Ratio of Insurer's Retention to Policy Limit (R/L)				
	0.00	0.25	0.50	0.75	0.95
10,000	0.80	0.95	1.02	1.07	1.10
100,000	0.89	1.24	1.36	1.44	1.50
500,000	0.92	1.31	1.40	1.47	1.52
1,000,000	0.93	1.31	1.39	1.47	1.53

table, undervalued premiums result from higher retention by the reinsured. This undervaluation is more important for higher policy limits. The greatest overvalued premiums are for large retention levels and policy limits. This is intuitively plausible; pricing in the tail of the distribution is very sensitive to model misspecification.

4.3.2 Estimating Regression Functions

As described in Section 1, the regression function is the most widely used tool for describing multivariate relationships. To illustrate in the context of losses and claims, we examine situations in which it is useful to estimate the expected amount of expenses for a given level of loss. Copulas can help us understand the full joint distribution and thus be used to address some important applications described in Sections 2, 4.3.1, and 5. We can also use copulas to define regression functions, as follows.

To be specific, let us assume an Archimedean form of the copula as in Section 3.3 so that the conditional distribution of X_k given $X_1, \ldots, X_{k-1}$ is

$$F_k\big(x_k | x_1, \ldots, x_{k-1}\big) = \frac{\phi^{-1(k-1)}\big\{c_{k-1} + \phi\big[F_k(x_k)\big]\big\}}{\phi^{-1(k-1)}\ (c_{k-1})},$$

where $c_k = \phi\{F_1(x_1)\} + \cdots + \phi\{F_k(x_k)\}$. Basic results from mathematical statistics show that the regression function can be expressed as

$$\begin{aligned} &E(X_k | x_1, \ldots, x_{k-1}) \\ &\quad = \int_0^\infty \big[1 - F_k(x | x_1, \ldots, x_{k-1})\big] dx \\ &\qquad + \int_{-\infty}^0 F_k(x | x_1, \ldots, x_{k-1})\, dx. \end{aligned}$$

In certain situations, this expression is convenient to evaluate. For example, assuming $k=2$ and using uniform marginals and Frank's copula, Genest (1987) gave the regression function

$$E\big(X_2 | X_1 = x\big) = \frac{(1 - e^{-\alpha})xe^{-\alpha x} + e^{-\alpha}\ (e^{-\alpha x} - 1)}{(e^{-\alpha x} - 1)(e^{-\alpha} - e^{-\alpha x})}.$$

In general, however, the calculation of the regression function is tedious. As an alternative, copulas are well-suited to the concept of "quantile" regression. Here, in lieu of examining the mean of a conditional distribution, one looks at the median or some other percentile (quantile) of the distribution. Specifically, define the p-th quantile to be the solution x_p of the equation:

$$p = F_k(x_p | x_1, \ldots, x_{k-1}).$$

For the case $k=2$, we have

$$p = F_2(x_p | X_1 = x_1) = C_1\big[F_1(x_1), F_2(x_p)\big]. \quad (4.5)$$

The first partial derivative for the case of the Gumbel-Hougaard copula is derived in Appendix A. For the case of Frank copula, the first partial derivative is given by

$$C_1(u, v) = \frac{e^{\alpha u}(e^{\alpha v} - 1)}{e^{\alpha} - 1 + (e^{\alpha u} - 1)(e^{\alpha v} - 1)}. \quad (4.6)$$

For a complete derivation of partial derivatives of the Frank's copula, see also Frees et al. (1996).

Thus, for a specified proportion p and amount of loss x_1, we can find the percentile of the corresponding expenses by solving Equation (4.5). For the case of the Gumbel-Hougaard copula, we use Equation (A.2) to get the p-th percentile of the expense level, which is given by

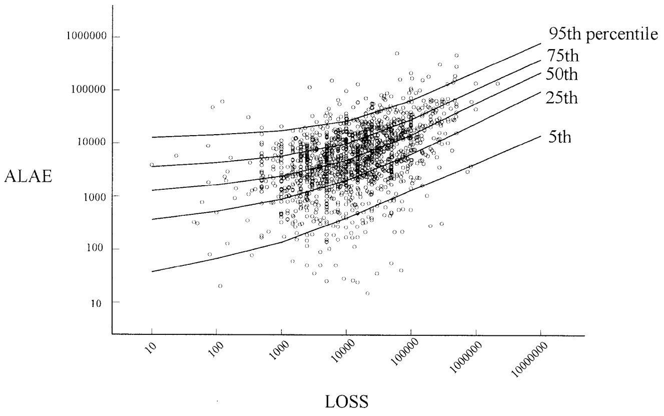
$$x_p = F_2^{-1}\ (v_p),$$

where v_p is the solution to the following equation:

$$\left(\frac{\ln F_1\ (x_1)}{\ln C\big(F_1(x_1),\ v_p\big)}\right)^{\alpha - 1} \frac{C\big(F_1(x_1),\ v_p\big)}{F_1(x_1)} = p.$$

For various percentiles, Figure 5 graphically displays the result of regression curves that provide estimates of expenses conditional on the amount of loss using the Gumbel-Hougaard copula. We superimposed these regression curves on the scatterplot of losses and expenses. This plot allows a manager to estimate expenses for a prespecified loss amount. By providing several percentiles, the manager can choose the degree of conservatism that is appropriate for the business decision at hand.

FIGURE 5
SCATTERPLOT OF ALAE VERSUS LOSS
WITH QUANTILE REGRESSION CURVES SUPERIMPOSED



5. ADDITIONAL APPLICATIONS OF COPULAS

We now discuss three different subject areas useful to the actuary in which copulas have been applied: stochastic ordering, fuzzy logic, and insurance pricing. Stochastic ordering refers to relationships among distribution functions of random variables. Fuzzy logic is an approach for dealing with uncertainty, analogous to probability theory. All actuaries are familiar with insurance pricing, which involves premium calculation.

5.1 Stochastic Ordering

In actuarial science and the economics of decision-making, the comparison of the attractiveness of various risks is usually of interest; therefore the subject of stochastic orderings is of prime importance. To illustrate, we say the random variable X_1 *stochastically dominates* the random variable X_2 with respect to a class of functions, say U, if for any function $u \in U$, we have

$$Eu(X_1) \geq Eu(X_2).$$

Often, u is the utility-of-wealth function that defines risk preferences of decision-makers. As special cases, we consider the well-known classes of first (FSD) and second (SSD) stochastic dominance. In FSD, the class U is taken to be the class of increasing utility functions. In SSD, U is the smaller class of increasing, concave utility functions, applicable to risk-averse decision-makers; see Kaas et al. (1994), Heilman and Schroter (1991), and Gooavaerts et al. (1982).

Copulas are used in the analysis for the demand of insurance coverage studied by Tibiletti (1995). The problem is to find the optimal insurance coverage for a decision-maker when a proportion of the wealth is uninsurable. Suppose that X_1 is the amount of uninsurable asset and X_2 is the insurable loss with $0 \leq x_2 \leq m$; that is, m is the maximum value of the insurable asset. In a two-period model, define the final wealth to be

$$Z = X_1 + m - X_2 + \delta(X_2 - p). \tag{5.1}$$

Here, δ is the coinsurance rate and p is the premium rate. In this insurance setup, the questions typically explored are: (1) what coinsurance rate δ maximizes $E(Z)$ and (2) when do certain "beneficial" changes that increase expected utility affect optimal coverage? Tibiletti explored beneficial changes such as (1) a shift in the distribution of either X_1 or X_2, (2) a change in the dependence between X_1 and X_2, and (3) a change in both (1) and (2). Because dependence comes into play, copulas provide a natural tool in this situation. Most of the theorems described and proved use conditions that involve a certain dependence ordering called *more concordance*. If (X_1, X_2) and (Y_1, Y_2) are pairs with associated copulas C_x and C_y, then the pair (X_1, X_2) is said to be *more concordant* than (Y_1, Y_2) if $C_x(a,b) \geq C_Y(a,b)$ for all $(a,b) \in [0,1]^2$.

The *more concordant* ordering is just one of several types of dependence ordering used to order multivariate distributions. See Tchen (1980), Kimeldorf and Sampson (1987, 1989), and Metry and Sampson (1991) for alternative orderings. For example, another type is the so-called *more regression dependent* ordering. Here, the random vector (X_1, X_2) is said to be *more regression dependent* than (Y_1, Y_2) if for any $x_1' > x_1$, we have

$$\frac{\Pr(X_2 \leq x_2 | X_1 = x_1')}{\Pr(Y_2 \leq y_2 | Y_1 = x_1')} \geq 1$$

whenever we have

$$\frac{\Pr(X_2 \leq x_2 | X_1 = x_1)}{\Pr(Y_2 \leq y_2 | Y_1 = x_1)} \geq 1.$$

See Bilodeau (1989). Dependence ordering is particularly useful for determining the range of possible dependence in multivariate random variables. Other areas in which stochastic orderings are useful are probability theory, reliability, and operations research; see Mosler and Scarsini (1991).

5.2 Fuzzy Sets

Like probability theory, fuzzy set theory deals with uncertainty. To start, let U be a non-empty set whose subsets are of primary interest; in fuzzy set theory, it is called the universe of discourse. A fuzzy set is defined to be a pair (E, μ_E) where $E \subset U$ and $\mu_E{:}U \rightarrow [0, 1]$. For any x in U, $\mu_E(x)$ denotes the degree to which x belongs to the fuzzy set, almost like a probability.

Operations, such as unions and intersections, on fuzzy sets are common. To illustrate, if we have two fuzzy sets (A, μ_A) and (B, μ_B), then their union is the fuzzy set (C, μ_C) where $\mu_C(x)=\max[\mu_A(x), \mu_B(x)]$ and their intersection is the fuzzy set (D, μ_D) with $\mu_D(x)=\min[\mu_A(x, \mu_B(x)]$. Other operations on fuzzy sets are performed using triangular norms. Triangular norms were considered originally in the context of probabilistic metric spaces and provide the link to copulas.

Following Ostaszewski (1993) and Klement (1982a, 1982b), a triangular norm (t-norm for short), T is a mapping $T{:}[0, 1]^2 \rightarrow [0, 1]$ with the following properties: (1) boundary condition: $T(x, 1)= x$; (2) monotonicity: $T(x, y) \leq T(u, v)$ whenever $x \leq u$, $y \leq v$; (3) commutativity: $T(x, y)=T(y, x)$; and (4) associativity: $T(T(x, y), z)=T(x, T(y, z))$. In Schweizer and Sklar (1983), several results are given that relate triangular norms and copulas. Many copulas can serve as triangular norms, as the following examples illustrate.

Example 5.1 Frechet Bounds

$T(x, y) = \min(x, y)$ and $T(x, y) = \max(x + y - 1, 0)$.

Example 5.2 Independence

$T(x, y) = xy$.

Example 5.3 Frank

$$T(x, y) = \frac{\ln 1 + \dfrac{(\eta^x - 1)(\eta^y - 1)}{\eta - 1}}{\ln \eta}$$

where $\eta > 0$, $\eta \neq 1$.

Therefore, by considering familiar copulas, as well as newly constructed copulas, we can define new operations that can be performed on fuzzy sets. Fuzzy set theory has been applied in specific areas of interest to many actuaries such as risk economics, interest theory, and underwriting/classification of risks. See Ostaszewski (1993), LeMaire (1990), and Young (1993) for descriptions of these applications.

5.3 Insurance Pricing

The calculation of a premium to be assigned to an insurance risk X is a fundamental job of the actuary. Recently, Wang (1996) developed a principle that exhibits several desirable properties of a premium principle. Wang (1997) used the concept of a distortion function and extended this, using copulas, to a portfolio of dependent risks.

Let X be an insurance risk with survival function $S(x)=\Pr(X>x)$. A class of premium principles can be defined by $\pi(X)=\int_0^\infty g[S(z)]dz$, where g, called the distortion function, is increasing with $g(0)=0$ and $g(1)=1$. In the special case of the function $g(t)=t^{1/\eta}$ with $\eta \geq 0$, we have the class of *proportional hazard* (*PH*) *transforms*.

For the bivariate case, consider a random vector (X_1, X_2) with distribution function $F(x_1, x_2)= C[F_1(x_1), F_2(x_2)]$; then the function $\widetilde{F}(x_1, x_2)=g[F(x_1, x_2)]$ is another joint distribution function with marginals $g(F_1)$ and $g(F_2)$. The associated copula is therefore $\widetilde{C}(u_1, u_2)=g[C(g^{-1}(u_1), g^{-1}(u_2))]$. Some illustrations of distortion functions follow.

Example 5.4 Power Distortion Function

Here, we have $g(t)=t^{1/\eta}$, where $\eta \geq 0$. The independence structure is preserved under this distortion function. To see this, note that if $C(u_1, u_2)=u_1u_2$, then $\widetilde{C}(u_1, u_2) =g[C(u_1^\eta, u_2^\eta)]=u_1u_2$.

Example 5.5 Exponential Distortion Function

Here, we have

$$g(t) = \frac{1 - e^{-\eta t}}{1 - e^{-\eta}},$$

where $\eta > 0$. Again, beginning with the independence copula $C(u_1, u_2)=u_1u_2$, it is straightforward to show that we can generate Frank's family of copulas

$$\widetilde{C}(u_1, u_2) = \ln \frac{1 + \dfrac{(\eta^{u_1} - 1)(\eta^{u_2} - 1)}{\eta - 1}}{\ln \eta}.$$

Using the concept of copulas, Wang (1997) extends the PH measure to a portfolio of risks. In many situations, for a portfolio of risks, the proposed measure is useful because individual risks are considered dependent. Wang justified dependence of individual risks by stating that they are "influenced by the same underlying market environment."

6. SUMMARY AND CONCLUSIONS

In analyzing the impact of future contingent events, actuaries are faced with problems involving multivariate outcomes. In this paper, we review the problems of (1) estimating distributions of joint lifetimes of paired individuals, useful in the analysis of survivorship insurance protection, and (2) investigating mortality experience, for the actuary who needs to distinguish among causes of death. We introduced, and provided a solution for, the problem of dependence between an insurer's losses and expenses. Failures of ignoring dependencies can lead to mispricing. Thus, it is important for actuaries to be able to adequately model multivariate outcomes.

The tool used to study multivariate outcomes is the copula function; it couples univariate marginals to the full multivariate distribution. The biological frailty models and the mathematical Archimedean models can motivate copulas. A statistical mixture of powers model serves as a bridge between these two sets of families.

Because copulas are parametric families, standard techniques such as maximum likelihood can be used for estimation. Other statistical tools have been recently developed to help fit copulas. We described a graphical tool to identify the form of the copula. We discussed how copulas could be used to simulate multivariate outcomes, an important tool for actuaries. We also developed the connection between copulas and the regression function, a widely used tool for summarizing what we expect based on conditional distributions.

This article has focused on the connection between copulas and statistics, the theory of data. Yet, much of the development of copulas has historically arisen from probability theory. To recognize this connection, we briefly reviewed topics in applied probability theory pertaining to copulas that are of the greatest interest to actuaries: stochastic ordering, fuzzy set theory, and insurance pricing.

Our knowledge of copulas has been rapidly developing recently. Many of the articles cited in this review paper were written in the 1990s. In another recent survey paper, Kotz (1997) cites three conferences in the last five years that were largely devoted to copulas. Copulas offer analysts an intuitively appealing structure, first for investigating univariate distributions and second for specifying a dependence structure. Copulas offer a flexible structure that can be applied in many situations. We hope that this article encourages actuaries to seek new applications for this promising tool.

ACKNOWLEDGMENTS

We thank Stuart Klugman for introducing the first author to the Section 4 insurance loss and expense problem at the 1995 Casualty Actuarial Society Ratemaking Seminar. Clive Keatinge of Insurance Services Office, Inc. provided the data for this problem and suggested the reinsurer's pricing example. We also thank Donald Jones, Marjorie Rosenberg, and Virginia Young for helpful comments from an early draft of this paper. The research of Emiliano Valdez was supported by a doctoral fellowship from the Society of Actuaries. The research of Edward Frees was partially supported by a professorship funded by the Time Insurance Company.

ANNOTATED BIBLIOGRAPHY*

ANDERSON, T.W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley.

BILODEAU, M. 1989. "On the Monotone Regression Dependence for Archimedean Bivariate Uniform," *Communications in Statistics–Theory & Methods A* 18:981–988.

This paper provides necessary and sufficient conditions for Archimedean copulas to exhibit the "monotone regression dependence" property. This property is desired for testing independence because of the resulting monotone power function. Frank's family and Cook and Johnson's family of copulas are shown to possess the property.

BOWERS, N., GERBER, H., HICKMAN, J., JONES, D., AND NESBITT, C. 1997. *Actuarial Mathematics*. 2nd ed. Schaumburg, Ill.: Society of Actuaries.

CARRIERE, J. 1994. "Dependent Decrement Theory," *Transactions of the Society of Actuaries* XLVI:45–74.

Copulas are used to model dependence in a multiple decrement framework. The problem of identifiability was briefly discussed. As an application, the paper uses U.S. population data to investigate the effect of removing heart/cerebrovascular diseases as a cause of death when assumed to be dependent with other causes.

CARRIERE, J. 1994. "A Large Sample Test for One-Parameter Families of Copulas," *Communications in Statistics–Theory and Methods* 23, no. 5:1311–1317.

This article provides an asymptotic test of whether or not identically and independently distributed bivariate data arise from a specified single-parameter copula family.

CLAYTON, D.G. 1978. "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of

*Articles directly related to copulas have been annotated for the reader's convenience.

Familial Tendency in Chronic Disease Incidence," *Biometrika* 65:141–151.

COMMITTEES ON ACTUARIAL PRINCIPLES OF THE SOCIETY OF ACTUARIES AND THE CASUALTY ACTUARIAL SOCIETY. "General Principles of Actuarial Science," April 30, 1997, discussion draft.

COOK, R.D., AND JOHNSON, M.E. 1981. "A Family of Distributions for Modeling Non-Elliptically Symmetric Multivariate Data," *Journal of the Royal Statistical Society* B 43:210–218.

The Cook and Johnson's family of copulas was first proposed in this paper. Designed for joint distributions that are not elliptically symmetric, the family includes, as special cases, the multivariate Burr, logistic, and Pareto distributions. Finally, the family was used to fit a uranium mining dataset.

COX, D.R., AND OAKES, D. 1984. *Analysis of Survival Data.* New York: Chapman and Hall.

DAVID, H.A., AND MOESCHBERGER, M.L. 1978. *The Theory of Competing Risks.* New York: MacMillan Publications.

ELANDT-JOHNSON, R.C., AND JOHNSON, N.L. 1980. *Survival Models and Data Analysis.* New York: John Wiley.

FRANK, M.J. 1979. "On the Simultaneous Associativity of F(x,y) and x+y−F(x,y)," *Aequationes Mathematicae* 19:194–226.

Frank's copula first appeared in this paper as a solution to a functional equation problem. That problem involved finding all continuous functions such that F(x,y) and x+y−F(x,y) are associative, defined in the paper. Associative functions are typically studies of concern in topological semigroups.

FREES, E. 1996. *Data Analysis Using Regression Models: The Business Perspective.* Englewood Cliffs: Prentice Hall.

FREES, E., CARRIERE, J., AND VALDEZ, E. 1996. "Annuity Valuation with Dependent Mortality," *Journal of Risk and Insurance* 63:229–261.

Copulas are used to value annuities based on more than one life in order to capture the possible dependence of lives. The study shows a reduction of approximately five percent in annuity values when dependent mortality models are used, compared to standard models of independence. The model was calibrated using data from a large insurer.

GALTON, F. 1885. "Regression Towards Mediocrity in Heredity Stature," *Journal of Anthropological Institute* 15:246–263.

GENEST, C. 1987. "Frank's Family of Bivariate Distributions," *Biometrika* 74:549–555.

Basic properties of the Frank's family of copulas are further investigated in this paper. Members of this family exhibit properties of stochastically ordered and likelihood ratio dependent distributions. Nonparametric estimators of the dependence parameter are recommended. Simulation procedures were used to evaluate performance of these estimators.

GENEST, C., AND MACKAY, J. 1986. "Copules Archimediennes et Familles de lois Bidimensionnelles Dont les Marges Sont Donnees," *The Canadian Journal of Statistics* 14:154–159.

This paper studies Archimedean copulas. Limits of sequences of members from this special class of copulas are investigated. Further, conditions for which these copulas exhibit properties of stochastically ordered distributions are provided. (Written in French.)

GENEST, C., AND MACKAY, J. 1986. "The Joy of Copulas: Bivariate Distributions with Uniform Marginals," *The American Statistician* 40:280–283.

The authors extend their previous work on Archimedean copulas by proving properties using basic calculus only. In particular, they prove a result that relates the corresponding Archimedean function to the familiar Kendall's tau measure of association.

GENEST, C., AND RIVEST, L. 1993. "Statistical Inference Procedures for Bivariate Archimedean Copulas," *Journal of the American Statistical Association* 88:1034–1043.

Assuming uniform marginals, this article introduces a procedure for estimating the function which determines an Archimedean copula. The estimate is fully nonparametric and is (pointwise) root-n consistent. Using a uranium exploration dataset, the article illustrates how to use the new procedure to choose a copula.

GENEST, C., GHOUDI, K., AND RIVEST, L. 1995. "A Semi-parametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions," *Biometrika* 82: 543–552.

GOOVAERTS, M.J., DE VYLDER, F., AND HAEZENDONCK, J. 1982. "Ordering of Risks: A Review," *Insurance: Mathematics and Economics* 1:131–163.

GOOVAERTS, M.J., DE VYLDER, F., AND HAEZENDONCK, J. 1984. *Insurance Premiums: Theory and Applications.* Amsterdam: North Holland.

GUMBEL, E.J. 1960. "Bivariate Exponential Distributions," *Journal of the American Statistical Association* 55:698–707.

HADAR, J., AND SEO, T.K. 1992. "A Note on Beneficial Changes in Random Variables," *The Geneva Papers on Risk and Insurance: Theory* 17:171–179.

HEILMAN, W.R., AND SCHROTER, K.J. 1991. "Ordering of Risks and their Actuarial Applications," in *Stochastic Orders and Decision under Risk,* edited by K. Mosler and M. Scarsini, Hayward, California: Institute of Mathematical Statistics.

HOUGAARD, P. 1984. "Life Table Methods for Heterogeneous Populations: Distributions Describing for Heterogeneity," *Biometrika* 71:75–83.

HOUGAARD, P. 1986. "A Class of Multivariate Failure Time Distributions," *Biometrika* 73:671–678.

HOUGAARD, P. 1987. "Modeling Multivariate Survival," *Scandinavian Journal of Statistics* 14:291–304.

This reviews models for multivariate survival and discusses practical situations for which they would be relevant. The author does not specifically suggest copula models for survival analysis but provides models based on marginal distributions. Here, the joint survival distribution is specified together with a single parameter to measure dependence. One can therefore infer that the joint survival can be expressed in terms of copulas. The paper is therefore appropriate for discussion of situations for which copula models are suitable.

Hougaard, P., Harvald, B., and Holm, N.V. 1992. "Measuring the Similarities Between the Lifetimes of Adult Danish Twins Born Between 1881–1930," *Journal of the American Statistical Association* 87:17–24.

Hutchinson, T.P., and Lai, C.D. 1990. *Continuous Bivariate Distributions, Emphasising Applications.* Adelaide, South Australia: Rumsby Scientific Publishing.

As the title indicates, this monograph organizes and describes the many families of bivariate distributions that have appeared in the literature. In addition to applications, special attention is given to copula families. In lieu of developing material from the foundations, the monograph collects a wide variety of properties and potential applications about many distributions, with a citation of original sources.

Jagger, C., and Sutton, C.J. 1991. "Death After Marital Bereavement—Is the Risk Increased?" *Statistics in Medicine* 10:395–404.

Joe, H. 1993. "Parametric Families of Multivariate Distributions with Given Margins," *Journal of Multivariate Analysis* 46: 262–282.

The author examines one-parameter bivariate families of copulas and describes some of their interesting properties. Whenever possible, these bivariate families were extended to the multivariate setting and corresponding properties studied. Several open problems are posed in the paper.

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

An excellent, although highly technical, research monograph that describes recent results in non-normal multivariate models. Much of the monograph focuses on copula constructions. It also contains discussion of categorical, longitudinal, and time series data.

Joe, H., and Hu, T. 1996. "Multivariate Distributions from Mixtures of Max-Infinitely Divisible Distributions," *Journal of Multivariate Analysis* 57:240–265.

This article explores the procedure based on the mixture of powers to generate new families of copulas. The procedure is based on distributions that are called max-infinitely divisible. Using this procedure, several families with closed-form multivariate copulas can be obtained, but only copulas that can be used for modeling positive dependence. It was mentioned in the discussion that extensions to allow for negative dependence will follow in a subsequent paper.

Johnson, N., and Kotz, S. 1972. *Distributions in Statistics: Continuous Multivariate Distributions.* New York: John Wiley.

Johnson, N., Kotz, S., and Balakrishnan, N. 1997. *Discrete Multivariate Distributions.* New York: John Wiley.

Johnson, M., and Tenenbein, A. 1981. "A Bivariate Distribution Family with Specified Marginals," *Journal of the American Statistical Association* 76:198–201.

This paper provides a procedure to construct families of bivariate distributions whose margins and measures of dependence are specified. The construction was based on the method of weighted linear combination. While not directly related to copulas, the procedure is well-suited for constructing bivariate families of copulas.

Johnson, R., and Wichern, D. 1988. *Applied Multivariate Statistical Analysis.* Englewood Cliffs, N.J.: Prentice Hall.

Kaas, R., Heerwaarden, A.E. van, and Goovaerts, M.J. 1994. *Ordering of Actuarial Risks.* Amsterdam: North Holland.

Kimeldorf, G., and Sampson, A.R. 1987. "Positive Dependence Orderings," *Annals of the Institute of Statistical Mathematics* 39:113–128.

Kimeldorf, G., and Sampson, A.R. 1989. "A Framework for Positive Dependence," *Annals of the Institute of Statistical Mathematics* 41:31–45.

Klement, E.P. 1982a. "Construction of Fuzzy σ-Algebras Using Triangular Norms," *Journal of Mathematical Analysis and Applications* 85:543–565.

Klement, E.P. 1982b. "Characterization of Fuzzy Measures Constructed by Means of Triangular Norms," *Journal of Mathematical Analysis and Applications* 86:345–358.

Klugman, S., and Parsa, A. 1995. "Fitting Bivariate Loss Distributions with Plackett's Model," Paper presented at the Casualty Actuarial Society Ratemaking Seminar.

Klugman, S., Panjer, H., Venter, G., and Willmot, G. 1997. *Loss Models: From Data to Decisions.* Unpublished Monograph.

Kotz, S. 1997. "Some Remarks on Copulas in Relation to Modern Multivariate Analysis," *Contemporary Multivariate Analysis and Its Applications* k.1–k.13.

Krzanowski, W.J. 1988. *Principles of Multivariate Analysis: A User's Perspective.* Oxford: Oxford University Press.

Lee, A.J. 1993. "Generating Random Binary Deviates Having Fixed Marginal Distributions and Specified Degrees of Association," *The American Statistician* 47:209–215.

Lemaire, J. 1990. "Fuzzy Insurance," *Astin Bulletin* 20:33–55.

Li, H., Scarsini, M., and Shaked, M. 1996. "Linkages: A Tool for the Construction of Multivariate Distributions with Given Nonoverlapping Multivariate Marginals," *Journal of Multivariate Analysis* 56:20–41.

Maguluri, G. 1993. "Semiparametric Estimation of Association in a Bivariate Survival Function," *Annals of Statistics* 21: 1648–1662.

This article considers bivariate survival data with no censoring of each risk. For the dependency between risks, a one-parameter copula model is used. With nonparametric estimates of the marginal survival function, this paper derives the asymptotic lower bound for the information in the copula parameter.

Makeham, W.M. 1874. "On an Application of the Theory of Decremental Forces," *Journal of the Institute of Actuaries* 18: 317–322.

Marshall, A.W., and Olkin, I. 1967. "A Multivariate Exponential Distribution," *Journal of the American Statistical Association* 62:30–44.

Marshall, A.W., and Olkin, I. 1988. "Families of Multivariate Distributions," *Journal of the American Statistical Association* 83:834–841.

Bivariate distributions with marginals as parameters can be generated using the procedure of mixtures. If F and G are univariate distribution functions, the distribution $H(x) = \int F^{\theta}(x)dG(\theta)$ can be generated. The paper studies properties of this mixture distribution and provides as examples familiar copulas such as Frank's, Cook and Johnson's, and Gumbel's. Extensions to the multivariate case are discussed.

Meester, S.G., and MacKay, J. 1994. "A Parametric Model for Cluster Correlated Categorical Data," *Biometrics* 50:954–963.

This article shows how to use copulas as a model for association in clustered or longitudinal data. The discussion of copulas is restricted to Frank's family, although extensions to other parametric families are indicated. Symmetric dependencies are considered, which is natural for clustered data. The responses are from the generalized linear model family and, in particular, may be categorical.

Metry, M.H., and Sampson, A.R. 1991. "A Family of Partial Orderings for Positive Dependence Among Fixed Marginal Bivariate Distributions," in *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*, edited by G. Dall'Aglio, S. Kotz, and G. Salinetti. The Netherlands: Kluwer Academic Publishers.

Meyer, J. 1992. "Beneficial Changes in Random Variables Under Multiple Sources of Risk and their Comparative Statics," *The Geneva Papers on Risk and Insurance: Theory* 17:7–19.

Mosler, K., and Scarsini, M. eds. 1991. *Stochastic Orders and Decision under Risk*. Hayward, California: Institute of Mathematical Sciences.

Nelsen, R.B. 1986. "Properties of a One-Parameter Family of Bivariate Distributions with Specified Marginals," *Communications in Statistics—Theory and Methods* 15: 3277–85.

This paper explores properties of Frank's one-parameter family of bivariate distributions. In particular, it offers formulas to evaluate three nonparametric measures of correlation, namely Spearman's rho, Kendall's tau, and the less familiar medial correlation coefficient. It also provides procedures to generate observations from this bivariate distribution, which may be used for simulation purposes.

Oakes, D. 1982. "A Model for Association in Bivariate Survival Data," *Journal of the Royal Statistical Society B* 44:414–422.

The author offers an alternative procedure to estimate the association parameter in the Cook and Johnson's family of bivariate distributions, which has been used to model association in bivariate survival data. The paper examines the likelihood procedure used to estimate the parameter by earlier authors. As an alternative, it proposes the use of nonparametric procedures using Kendall's coefficient. The asymptotic variance of this estimator is evaluated.

Oakes, D. 1989. "Bivariate Survival Models Induced by Frailties," *Journal of the American Statistical Association* 84: 487–493.

This article shows how Archimedean copulas arise naturally in the context of frailty models. The work emphasizes a special case of Marshall and Olkin (1989): bivariate survival times with a common latent mixing parameter. The paper introduces a "cross-ratio" function that provides a useful diagnostic tool for identifying the survival distribution based on bivariate survival data.

Oakes, D. 1994. "Multivariate Survival Distributions," *Journal of Nonparametric Statistics*, 343–354.

This article reviews properties of latent mixtures of survival distributions that give rise to frailty models.

Ostaszewski, K. 1993. *Fuzzy Set Methods in Actuarial Science.* Schaumburg: Society of Actuaries.

Parkes, C.M., Benjamin, B., and Fitzgerald, R.G. 1969. "Broken Heart: A Statistical Study of Increased Mortality Among Widowers," *British Medical Journal* 1:740–743.

Scarsini, M. 1984. "On Measures of Concordance," *Stochastica* 8:201–219.

Schweizer, B. 1991. "Thirty Years of Copulas," in *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*, edited by G. Dall'Aglio, S. Kotz, and G. Salinetti. The Netherlands: Kluwer Academic Publishers.

This article reviews the early development of copulas. The historical overview focuses on beginnings with a focus on probability theory. Properties of n-dimensional distribution functions are reviewed and their extensions to probabilistic metric spaces are discussed. Relationships to triangular norming functions are developed that provide bounds on functions of distribution functions. The paper also gives a useful introduction to the role of copulas for developing measures of dependence and their relationship with the dependence axioms of Renyi.

Schweizer, B., and Sklar, A. 1983. *Probabilistic Metric Spaces.* New York: North Holland.

A probabilistic metric space is a set, or space, of distribution functions combined with a function, or metric, that measures the distance between two elements of the set. This monograph develops probabilistic metric spaces. A special feature is that an entire chapter (six) is devoted to copulas. Further, the relationship between copulas and triangular norms is emphasized.

Schweizer, B., and Wolff, E.F. 1981. "On Nonparametric Measures of Dependence for Random Variables," *The Annals of Statistics* 9:879–885.

This paper uses copulas to define three nonparametric measures of dependence for pairs of random variables. In particular, the familiar measures of dependence Pearson's correlation, Spearman's rho, and Kendall's tau are expressed in terms of copulas.

Seal, H.L. 1967. "Studies in the History of Probability and Statistics XV: The Historical Development of the Gauss Linear Model," *Biometrika* 54:1–24.

Seal, H.L. 1977. "Multiple Decrements or Competing Risks," *Biometrika* 64:429–439.

Shih, J.H., and Louis, T.A. 1995. "Inferences on the Association Parameter in Copula Models for Bivariate Survival Data," *Biometrics* 51:1384–1399.

This paper studies bivariate survival data where the risks may be dependent but do not preclude one another. Each of the two risks is subject to censoring by an independent mechanism. The dependency between risks is modeled using a one-parameter copula function. The marginal survival functions are estimated (i) nonparametrically, using Kaplan-Meier estimators and (ii) parametrically. A two-stage estimation procedure is proposed; both finite and asymptotic properties are established. An AIDS dataset illustrates the procedures.

Sklar, A. 1959. "Fonctions de repartition a n dimensions et leurs marges," *Publ. Inst. Stat. Univ. Paris* 8:229–231.

Sklar, A. 1973. "Random Variables, Joint Distribution Functions and Copulas," *Kybernetika* 9:449–460.

The author is one of the pioneers of copulas. In this paper, he proves elementary results that relate copulas to distribution functions and random variables. In particular, given a copula function C and an n-tuple distribution function $(F_1, \ldots, F_n)$, he establishes the existence of a probability space and a set of random variables $X_1, \ldots, X_n$ defined over that space with C as the associated copula. In addition, the paper describes structural properties of a special type of copula called associative copulas.

Stigler, S.M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Mass.: Harvard University Press.

Sungur, E.A. 1990. "Dependence Information in Parameterized Copulas," *Communications in Statistics: Simulations* 19: 1339–1360.

This paper provides mechanisms for extracting the dependence information from parameterized copulas. The author defines various dependence functions to capture dependence information. Approximations to copulas using Taylor's expansion are discussed. Most of the derivations relate to two-dimensional copulas although a brief attempt to extend to three-dimensional copulas is made.

Tchen, A. 1980. "Inequalities for Distributions with Given Marginals," *Annals of Probability* 8:814–827.

Tibiletti, L. 1995. "Beneficial Changes in Random Variables via Copulas: An Application to Insurance," *The Geneva Papers on Risk and Insurance: Theory* 20:191–202.

Vaupel, J.W., Manton, K.G. and Stallard, E. 1979. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality," *Demography* 16:439–454.

Wang, S. 1996. "Premium Calculation by Transforming the Layer Premium Density," *ASTIN Bulletin* 26:71–92.

Wang, S. 1997. "PH-Measure of Portfolio Risks," Chapter 9 of unpublished lecture notes, *Risk Measures with Applications in Insurance Ratemaking and Actuarial Valuation*. Waterloo, Ont.: University of Waterloo.

Ward, A. 1976. "Mortality of Bereavement," *British Medical Journal* 1:700–702.

Young, V. 1993. "The Application of Fuzzy Sets to Group Health Underwriting," *Transactions of the Society of Actuaries* XLV:551–590.

Zheng, M., and Klein, J.P. 1995. "Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula," *Biometrika* 82:127–138.

For bivariate survival data, it is well-known that without assumptions about the relationship between the competing survival times, the marginal survival functions are unidentifiable. This article assumes a known copula and introduces nonparametric estimators of the marginal survival functions. The estimates reduce to the Kaplan-Meier product limit estimator in the case of independent survival times. Further, bounds on the survival functions are provided when the copula is known only approximately.

APPENDIX A

GUMBEL-HOUGAARD COPULA AND ITS PARTIAL DERIVATIVES

In this appendix, we derive the formulas needed to evaluate the likelihood in Equation (4.3) in the case in which we apply the Gumbel-Hougaard copula as given in Table 1. First, we re-express the Gumbel-Hougaard copula as follows:

$$\left[-\ln C(u, v)\right]^{\alpha} = (-\ln u)^{\alpha} + (-\ln v)^{\alpha}.$$

For simplicity, we denote $C(u, v)$ by C. Now, take the partial derivative with respect to u of both sides of the equation to get:

$$\frac{(-\ln C)^{\alpha-1}}{C}\frac{\partial C}{\partial u} = \frac{(-\ln u)^{\alpha-1}}{u}. \qquad (A.1)$$

Solving for $\partial C/\partial u$, we have:

$$\frac{\partial C}{\partial u} = \left(\frac{\ln u}{\ln C}\right)^{\alpha-1}\frac{C}{u}. \qquad (A.2)$$

Applying symmetry, we get:

$$\frac{\partial C}{\partial v} = \left(\frac{\ln v}{\ln C}\right)^{\alpha-1}\frac{C}{v}. \qquad (A.3)$$

From Equation (A.1), we can take the partial derivative with respect to v to get:

$$\frac{(-\ln C)^{\alpha-1}}{C}\frac{\partial^2 C}{\partial u \partial v} - \frac{\partial C}{\partial u}\frac{\partial C}{\partial v}\frac{(-\ln C)^{\alpha-1}}{C^2}\left[(\alpha-1)(-\ln C)^{-1}+1\right] = 0.$$

Rearranging terms yields the second partial derivative of the Gumbel-Hougaard copula:

$$\frac{\partial^2 C}{\partial u \partial v} = \frac{1}{C}\frac{\partial C}{\partial u}\frac{\partial C}{\partial v}\left[(\alpha-1)(-\ln C)^{-1}+1\right]. \quad \text{(A.4)}$$

Equations (A.2), (A.3), and (A.4) are used to evaluate the log-likelihood given in Equation (4.3).

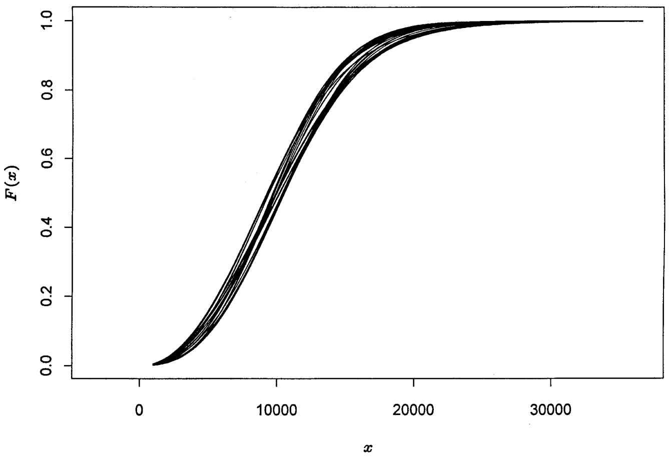
Appendix B

Table of One-Parameter Family of Copulas

Family	General Form of Copula	Parameter Constraint	Kendall's Tau	Spearman's Rho
Ali-Mikhail-Haq	$uv[1-\alpha(1-u)(1-v)]^{-1}$	$-1 \le \alpha \le 1$	$\left(\frac{3\alpha-2}{\alpha}\right) - \frac{2}{3}\left(1-\frac{1}{\alpha}\right)^2 \ln(1-\alpha)$	Complicated form
Cook-Johnson	$[u^{-\alpha}+v^{-\alpha}-1]^{-1/\alpha}$	$\alpha \ge 0$	$\frac{\alpha}{\alpha+2}$	Complicated form
Farlie-Gumbel-Morgenstern	$uv[1+\alpha(1-u)(1-v)]$	$-1 \le \alpha \le 1$	$\frac{2}{9}\alpha$	$\frac{1}{3}\alpha$
Frank	$\frac{1}{\alpha}\ln\left[1+\frac{(e^{\alpha u}-1)(e^{\alpha v}-1)}{(e^{\alpha}-1)}\right]$	$\alpha \ne 0$	$1-\frac{4}{\alpha}[D_1(-\alpha)-1]$	$1-\frac{12}{\alpha}[D_2(-\alpha)-D_1(-\alpha)]$
Gumbel-Hougaard	$\exp\left\{-[(-\ln u)^{\alpha}+(-\ln v)^{\alpha}]^{1/\alpha}\right\}$	$\alpha \ge 1$	$1-\alpha^{-1}$	No closed form
Normal (Bivariate)	$H\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$ where H is the bivariate normal distribution function with correlation coefficient α and Φ^{-1} is the inverse of a univariate normal distribution function.	$-1 \le \alpha \le 1$	$\frac{2}{\pi}\arcsin(\alpha)$	$\frac{6}{\pi}\arcsin\left(\frac{\alpha}{2}\right)$
Plackett	$\frac{1}{2}(\alpha-1)^{-1}\left\{1+(\alpha-1)(u+v)-\left[\left(1+(\alpha-1)(u+v)\right)^2+4\alpha(1-\alpha)\right]^{1/2}\right\}$	$\alpha \ge 0$	No closed form	$\frac{(\alpha+1)}{(\alpha-1)} - \frac{2\alpha\ln\alpha}{(\alpha-1)^2}$

2. The GB2 model, however, is not suitable for pricing or for reinsurance since it gives a lower value of x than the gamma model for almost all values of p.
3. The mixture of two-gamma models, after applying the lemma below, becomes a four-parameter model. Its log-likelihood is marginally better than that of GB2, and we expect its confidence region is about the same width as the GB2.

Figure 4
***F*(*x*|*a*, *b*, *p*, *q*) with (*a*, *b*, *p*, *q*) Running through Its 95% Confidence Region**



Mixture of Linear Exponential Objects

For both normal distribution and gamma distribution, the mean of the MLE-fitted model must match the empirical mean. For the normal, this result is found in any Course 110 textbook. For the gamma, or mixture of gamma and normal, it's not difficult to prove.

Lemma (Mixture of Exponential Family with a Linear Term)

If X has the density function

$$f(x) = h(\alpha, x)c(\alpha, \beta)\exp(-\beta x), \tag{1}$$

then

$$E_f(X) = \int xf(x)\,dx = \frac{\dfrac{\partial c}{\partial \beta}(\alpha, \beta)}{c(\alpha, \beta)} \tag{2}$$

and the mean of the maximum likelihood estimated density function of the form (1) fitted to observed data would match the empirical mean:

$$E_f(X) = \frac{\dfrac{\partial c}{\partial \beta}(\hat{\alpha}, \hat{\beta})}{c(\hat{\alpha}, \hat{\beta})} = \frac{\sum x}{n}. \tag{3}$$

And a similar result holds if X is a mixture of two densities satisfying Equation (1)

$$\begin{aligned} f(x) &= pf_1(x) + (1 - p)f_2(x) \\ &= ph_1(\alpha_1, x)c_1(\alpha_1, \beta_1)\exp(-\beta_1 x) \\ &\quad + (1 - p)h_2(\alpha_2, x)c_2(\alpha_2, \beta_2)\exp(-\beta_2 x). \end{aligned} \tag{4}$$

Then the formula mean of the maximum likelihood estimated density function of the form (4) fitted to observed data $\{x_1, \ldots, x_n\}$ would match the empirical mean:

$$E_f(X) = \frac{\dfrac{\partial c}{\partial \beta}(\alpha, \beta)}{c(\alpha, \beta)} = \frac{\sum x}{n}.$$

This lemma says, for example, that a gamma mixed with normal fitted by MLE will match the empirical mean.

REFERENCE

COXETER, H.S.M. 1973. *Regular Polytopes,* 3rd ed. New York: Dover.

"Understanding Relationships Using Copulas," by Edward Frees and Emiliano Valdez, January 1998

CHRISTIAN GENEST,* KILANI GHOUDI†, AND LOUIS-PAUL RIVEST‡

This article aimed to introduce actuaries to "copulas"—that is, distributions whose univariate marginals are uniform—as a tool for understanding relationships among multivariate outcomes. Through its limpid exposition of some of the recent developments in

*Christian Genest, Ph.D., is Professor of Statistics, Département de mathématiques et de statistique, Université Laval, Sainte-Foy, Québec, Canada G1K 7P4, e-mail, genest@mat.ulaval.ca.

†Kilani Ghoudi, Ph.D., is Professor of Statistics, Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada G9A 5H7, e-mail, ghoudi@uqtr.uquebec.ca.

‡Louis-Paul Rivest, Ph.D., is Professor of Statistics, Département de mathématiques et de statistique, Université Laval, Sainte-Foy, Québec, Canada G1K 7P4, e-mail, lpr@mat.ulaval.ca.

statistical modeling with copulas, this work and its substantial annotated bibliography should contribute much to the dissemination of these techniques in actuarial circles.

In their paper, the authors explain how it is possible, using copulas, to dissociate the choice of marginal distributions from that of a model describing the dependence between pairs of correlated variables such as the economic loss and the medical indemnity component of a disability policy. Part of their discussion revolves around the issue of identifying an appropriate family of copulas and the estimation of its parameters from a multivariate random sample. In Section 4.2.1, they show specifically how an indemnity payment X_1 (loss) and an allocated loss adjustment expense (ALAE) X_2 could profitably be modeled from a random sample of 1,500 liability claims. The copula they select is of the Archimedean variety, which means that for all $0 \le u, v \le 1$, it may be written in the form

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

for some decreasing, convex function ϕ: $(0,1] \rightarrow [0, \infty)$ such that $\phi(1) = 0$, with the convention that $\phi^{-1}(t) = 0$ for all $t \ge \lim_{s \downarrow 0} \phi(s) \equiv \phi(0)$. In that context, they find that the generator $\phi(t) = \log^{\alpha}(1/t)$ of Gumbel's family of copulas provides an adequate fit, given a suitable choice of parameter $\alpha \ge 1$.

This discussion expands on the issue of selecting an appropriate copula for modeling purposes. Cast in terms of the above example, the question to be addressed is whether it is possible to improve on Gumbel's family as a model for describing the relationship between variables loss and ALAE. Techniques for imbedding copulas of the form (1) in larger models of exchangeable and nonexchangeable variables are introduced, and the dataset considered by Frees and Valdez is used to illustrate how such extensions can yield improvements in the fit of a model. Although the proposed methodology does not lead to a better solution in this particular application, it should prove useful in a variety of contexts. Finally, the attention of the actuarial community is directed to another potentially useful collection of bivariate copulas whereof Gumbel's family is a distinguished subset: the class of extreme value copulas.

Generating Archimedean Copula Models

As pointed out by Frees and Valdez, and several others before them, many well-known systems of bivariate distributions have underlying copulas of the form (1). These Archimedean copulas (Genest and MacKay 1986) provide a host of models that are versatile, in terms of both the nature and strength of the association they induce between the variables. It is not surprising therefore that they have been used successfully in a number of data-modeling contexts, especially in connection with the notion of "frailty" (Clayton 1978, Oakes 1989, Zheng and Klein 1995, Bandeen-Roche and Liang 1996, Day, Bryant and Lefkopoulou 1997).

Because copulas characterize the dependence structure of a random vector once the effect of the marginals has been factored out, identifying and fitting a copula to data poses special difficulties. The families of Clayton (1978), Gumbel (1960), and Frank (1979), for example, provide handy, one-parameter representations of the association between variables X_i with marginal distribution functions F_i. However, they are unsuitable in those situations in which the joint dependence structure of the uniform random variables $F_i(X_i)$ is not exchangeable, because Archimedean copulas in general are symmetric in their arguments. Unfortunately, diagnostic procedures that can help delineate circumstances where model (1) is adequate have yet to be analyzed.

Starting from the assumption that the Archimedean dependence structure is appropriate in a bivariate context, Frees and Valdez explain how the choice of the generator can be made, using the technique developed by Genest and Rivest (1993). Given observations from a random pair (X_1, X_2) with distribution H, this procedure relies on the estimation of the univariate distribution function $K(v)$ associated with the "probability integral transformation," $V = H(X_1, X_2)$. For each parametric Archimedean structure C_γ that is envisaged, an estimator $\hat{\gamma}$ of γ is obtained, and the $K_{\hat{\gamma}}$'s are compared, graphically or otherwise, to a nonparametric estimator of K, whose stochastic behavior as a process has recently been studied by Barbe, Genest, Ghoudi and Rémillard (1996).

Having identified an Archimedean family of copulas that provides a reasonable fit to a bivariate random sample, are there simple ways of generating alternative models that might improve this fit? One suggestion consists of enlarging the selected family through various combinations of the following rules, which can be used to construct nested classes of Archimedean generators. For ease of reference, this collection of rules is presented as a proposition, whose proof is straightforward and left to the reader.

Proposition 1

Suppose that ϕ is the generator of a bivariate Archimedean copula. In other words, assume that ϕ: $(0,1]$

→ [0, ∞) is a decreasing, convex function such that $\phi(1) = 0$.

(i) (*right composition*) If f: [0, 1] → [0, 1] is an increasing, concave bijection, then $\phi \circ f(t) = \phi\{f(t)\}$ generates a bivariate Archimedean copula.

(ii) (*left composition*) If f: [0, ∞) → [0, ∞) is an increasing, convex function such that $f(0) = 0$, then $f \circ \phi$ generates a bivariate Archimedean copula.

(iii) (*scaling*) If $0 < \alpha < 1$, then $\phi_\alpha(t) = \phi(\alpha t) - \phi(\alpha)$ generates a bivariate Archimedean copula.

(iv) (*composition via exponentiation*) If $(\phi')^2 \leq \phi''$ and if ψ generates a bivariate Archimedean copula, then $\psi(e^{-\phi})$ also generates a bivariate Archimedean copula.

(v) (*linear combination*) If α and β are positive reals and if ψ generates a bivariate Archimedean copula, then $\alpha\psi + \beta\phi$ also generates a bivariate Archimedean copula.

To illustrate these composition rules, define $\phi(t) = \log(1/t)$ for all $0 < t \leq 1$ and consider the function $f_\alpha(t) = (e^{\alpha t} - 1)/\alpha$ on [0, ∞), which satisfies the requirements of part (ii) of the proposition for arbitrary $\alpha > 0$. Then

$$\phi_\alpha(t) = f_\alpha\{\phi(t)\} = (t^{-\alpha} - 1)/\alpha, \qquad 0 < t \leq 1$$

is immediately recognized as the generator of Clayton's family of bivariate copulas. Frank's and many other classical systems of bivariate Archimedean distributions can be recovered in this fashion, often with the generator $\log(1/t)$ of the independence copula as a starting point. Special cases of rules (i) and (ii) can be found in the work of Oakes (1994) and Nelsen (1997). For some examples of Archimedean (as well as non-Archimedean) copulas and for additional ways of combining them to build new bivariate and multivariate distributions, see for example Joe (1993) or Joe and Hu (1996).

The usefulness of Proposition 1 is best demonstrated in a data analytic context. In their paper, for example, Frees and Valdez envisage Clayton's, Frank's, and Gumbel's bivariate copulas for modeling the loss and expense components of an insurance company's indemnity claims. As explained below, a specific combination of composition rules (i) and (iii) makes it possible to include these three families as special cases of a single system of bivariate Archimedean copulas, whose generator is given by

$$\phi_{\alpha,\beta,\gamma}(t) = \log\left\{\frac{1 - (1 - \gamma)^\beta}{1 - (1 - \gamma t^\alpha)^\beta}\right\}, \qquad 0 < t \leq 1 \tag{2}$$

with arbitrary $\alpha > 0$, $\beta > 1$ and $0 < \gamma < 1$.

The interest of this enlarged model for inferential purposes is twofold. On one hand, it may sometimes provide a significantly better fit of the data than any submodel. On the other, it suggests a simple statistical procedure for choosing between the one-parameter models: only that which is "closest" to the full model, in some sense, need be chosen. Tests of significance of the appropriate parameters can be used to assist in this choice.

A similar strategy can be designed to handle situations in which nonexchangeability is suspected between the $F_i(X_i)$'s. It is briefly described in the next section, and a concrete application is given in the following section using the liability data of Frees and Valdez. The following paragraphs substantiate some of the above claims concerning the three-parameter, bivariate Archimedean family of copulas generated by (2).

To check that $\phi_{\alpha,\beta,\gamma}$ is a valid generator of bivariate Archimedean copulas, start from $\phi(t) = \log(1/t)$, $0 < t \leq 1$, and use the fact that $f_\beta(t) = 1 - (1 - t)^\beta$ is increasing and concave on [0, 1] for $\beta > 1$ to conclude from (i) that $\phi_1(t) = -\log\{1 - (1 - t)^\beta\}$ generates a bivariate Archimedean copula. Next, apply rule (iii) to see that $\phi_2(t) = \phi_1(\gamma t) - \phi_1(\gamma)$ also generates a bivariate Archimedean copula. Finally, observe that since t^α is a concave, increasing function on [0, 1] for $0 < \alpha < 1$, $\phi_{\alpha,\beta,\gamma}(t) = \phi_2(t^\alpha)$ is indeed a bivariate Archimedean copula generator because of rule (i). Note that function (2) is actually convex and decreasing even when $\alpha \geq 1$.

The parameter values of (2) that yield the Clayton, Frank and Gumbel families of copulas are given in Table 1, in which the notation $o(x)$ stands for a function $f(x)$ satisfying $f(x)/x \to 0$ as $x \to 0$. Note in particular that Clayton's and Gumbel's families of bivariate copulas obtain when α and $1 - \gamma$ are both going to zero. If α converges to zero faster than $1 - \gamma$, Clayton's system is actually the right limit, while when $1 - \gamma$ is going to zero faster than α, the limit is a Gumbel copula.

Table 1
Parameter Values of Archimedean Generator (2) Yielding Clayton's, Frank's, and Gumbel's Families with Dependence Parameter δ

Family	α	β	γ
Clayton	$o(1-\gamma)$	$(1-\gamma)\delta/\alpha$	$\uparrow 1$
Frank	1	$\uparrow \infty$	$-\delta/\beta$
Gumbel	$\downarrow 0$	δ	$1-o(\alpha)$

Generating Asymmetric Copulas

All bivariate copulas C in the Archimedean class satisfy the exchangeability condition $C(u, v) = C(v, u)$ on their domain. If a family of copulas of form (1) is envisaged as a model in a situation in which the appropriateness of this symmetry condition is doubtful, one may wish to enlarge the system to include nonexchangeable models. The following proposition shows how this can be done and paves the way to a simple test of nonexchangeability, which is illustrated in the next section.

Proposition 2

Let C be an exchangeable bivariate copula. A family of nonexchangeable bivariate copulas $C_{\kappa,\lambda}$ with parameters $0 < \kappa, \lambda < 1$ that includes C as a limiting case is defined by

$$C_{\kappa,\lambda}(u, v) = u^{1-\kappa}v^{1-\lambda}C(u^{\kappa}, v^{\lambda}), \qquad 0 \le u, v \le 1.$$

This mechanism for generating asymmetric copulas was first studied by Khoudraji (1995) in an unpublished doctoral dissertation prepared under the supervision of the first and third authors. An interesting property of this nonexchangeable model is that it is easy to generate random variates distributed according to $C_{\kappa,\lambda}$. Indeed, if the pair (U_1, V_1) is drawn from copula $C(u, v)$ and if U_2 and V_2 are independent observations from a uniform distribution on the interval $[0, 1]$, then $C_{\kappa,\lambda}(u, v)$ is the joint distribution of

$$U = \max\{U_1^{1/\kappa}, U_2^{1/(1-\kappa)}\}, \; V = \max\{V_1^{1/\lambda}, V_2^{1/(1-\lambda)}\}.$$

The verification of this assertion also constitutes a proof that the function $C_{\kappa,\lambda}$ defined in Proposition 2 is always a bivariate copula. A natural extension of this result is the case in which the pair (U_2, V_2) is distributed as a copula $D(u, v)$, which may be different from independence. The details are given in Chapter 4 of Khoudraji (1995).

Seeking Improvements in the Fit of a Model

To illustrate concretely how Propositions 1 and 2 can help enhance the fit provided by an Archimedean copula model, it is convenient to reconsider the loss and ALAE data of Frees and Valdez. Specifically, this section investigates whether a better model than Gumbel's can be found for these data, either using copula

$$C(u, v) = \phi^{-1}_{\alpha,\beta,\gamma}\{\phi_{\alpha,\beta,\gamma}(u) + \phi_{\alpha,\beta,\gamma}(v)\} \qquad (3)$$

or an asymmetric copula constructed by the method described in the previous section.

In their paper, Frees and Valdez explain how maximum-likelihood-based computer algorithms can assist in estimating simultaneously the parameters associated with the marginal distributions of a random pair (X_1, X_2) and those which correspond to the family of copulas selected as a model for dependence. Applying this procedure to the three-parameter copula (3) and Pareto marginals, say, would thus require the estimation of seven parameters altogether, nine if the copula were also expanded via Proposition 2 to check for asymmetry.

While this approach is sound and straightforward to implement, it might yield inappropriate values of the dependence parameters α, β and γ (and eventually κ and λ) if the parametric models chosen for X_1 and X_2 turned out to be incorrect. There is no serious reason to doubt this choice in the present case, but as a general precaution, one may wish to ensure that uncertainty about the marginals does not affect unduly the parameter estimates of the copula model. In other words, a parametric estimation procedure that is robust to the choice of marginal distributions is called for.

A margin-free parameter estimation procedure for copulas has been described in broad, nontechnical terms by Oakes (1994). This semiparametric technique was subsequently developed and studied in Genest, Ghoudi and Rivest (1995) and adapted to the case of censoring by Shih and Louis (1995). In that pseudo-likelihood-based approach, the contribution of an uncensored individual bivariate data point to the likelihood is, in the notation of Frees and Valdez,

$$C_{12}\{F_{1n}(x_1), F_{2n}(x_2)\},$$

where C_{12} is the mixed partial derivative of the specified parametric copula, and $F_{in}(x_i)$ stands for the empirical distribution function of X_i, $i = 1, 2$. In the application at hand, $F_{1n}(x)$ and $F_{2n}(x)$ would thus be the proportions of claims for which loss X_1 and ALAE X_2 are less than or equal to x, respectively.

When the copula C considered involves three parameters, as in model (3), computing the mixed partial derivative C_{12} and maximizing the pseudo-likelihood may seem like an insurmountable task. Fortunately, there is no need to have a closed analytical expression for C_{12} to produce numerical estimates of the dependence parameters. Using a software for symbolic manipulations, such as Maple, one can feed in $\phi_{\alpha,\beta,\gamma}$ and its inverse and get, as an output, subroutines for the numerical evaluation of C and C_{12}. These subroutines can then be linked to a general-purpose maximization program that finds the pseudo-likelihood estimates.

The calculations presented below were obtained with a Fortran program developed by S. G. Nash from George Mason University. Formulas for calculating the standard errors of the estimates are given in Genest, Ghoudi and Rivest (1995). The slight censoring in X_1 was ignored in the calculations, because the numerical results presented by Frees and Valdez suggest that this has a negligible impact on the estimates. Alternatively, the pseudo-likelihood procedure of Shih and Louis (1995) could be implemented to account for the censoring in X_1.

The semiparametric parameter estimates for Clayton's, Frank's, and Gumbel's families of copulas are given in Table 2, along with those of the three parameters of the Archimedean model generated by (2). The estimates obtained for Frank's and Gumbel's models are very close to those reported by Frees and Valdez (1998). This provides indirect evidence that their choice of Pareto distributions for the marginals is adequate.

As explained earlier, the graphical technique suggested by Genest and Rivest (1993) can be used to investigate the fit of the four Archimedean models at hand. However, the plots obtained with parameters estimated from the pseudo-likelihood tend to be more sensitive to model misspecification than those constructed with estimates derived from Kendall's tau. Alternatively, the value of the maximized pseudo-log-likelihood may be used, formally or informally, to judge the models' relative merits.

In Table 2, $1 - \hat{\gamma}$ is much smaller than $\hat{\alpha}$, and neither parameter is significantly different from 0. As mentioned earlier, this is a situation in which Gumbel's family is indicated. The small difference observed between the maximized pseudo-log-likelihoods for the three-parameter model and Gumbel's submodel confirms that the latter fits well.

Using Gumbel's copula as a starting point, we may also wish to investigate the issue of asymmetry. This can be done easily with the technique presented above. The maximum pseudo-log-likelihood is 207.3 and the parameter estimates are $\hat{\alpha} = 1.47$ (s. e. = 0.07), $\hat{\kappa} = 0.94$ (s. e. = 0.09), $\hat{\lambda} = 1$. Because $\hat{\kappa}$ is not significantly different from 1, it appears that in this case, the introduction of asymmetry in Gumbel's copula does not improve the fit.

Extreme Value Copulas

Whenever Gumbel's family of copulas seems adequate for a bivariate dataset, one may also wish to check whether a better fit could be achieved by another system in the maximum extreme value class. These are copulas that can be expressed as

$$C_A(u, v) = \exp\left[\log(uv) A \left\{\frac{\log(u)}{\log(uv)}\right\}\right], \qquad 0 \le u, v \le 1 \quad (4)$$

in terms of a dependence function A: $[0, 1] \rightarrow [1/2, 1]$, which is convex and verifies $A(t) \ge \max(t, 1 - t)$ for all $0 \le t \le 1$. It is easy to see that the choice

$$A(t) = \{t^\alpha + (1 - t)^\alpha\}^{1/\alpha}, \qquad 0 \le t \le 1$$

corresponds to Gumbel's family of copulas, which is the only one that can be written simultaneously in the

Table 2
Maximized Pseudo-log-likelihood and Parameter Estimates, with Their Standard Errors, Associated with Four Archimedean Copula Models

Family	Pseudo-Log-Likelihood	$\hat{\alpha}$ (s. e.)	$\hat{\beta}$ (s. e.)	$\hat{\gamma}$ (s. e.)
Clayton	93.8	0.52 (0.032)	—	—
Frank	172.5	−3.10 (0.57)	—	—
Gumbel	207.0	1.44 (0.032)	—	—
Three-Parameter	210.1	0.138 (0.147)	1.55 (0.92)	0.9986 (0.0025)

forms (1) and (4) (Genest and Rivest 1989). Additional parametric families of maximum extreme value copulas are given by Tawn (1988) and Anderson and Nadarajah (1993), among others.

The terminology for model (4) can be justified as follows. Let $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ be a random sample from bivariate distribution H, define $M_{in} = \max(X_{i1}, \ldots, X_{in})$, $i = 1, 2$, and suppose that there exist constants $a_{in} > 0$ and b_{in} for which the pair

$$\left(\frac{M_{1n} - b_{1n}}{a_{1n}}, \frac{M_{2n} - b_{2n}}{a_{2n}}\right)$$

has a non-degenerate, joint limiting distribution H^*. The marginal distributions of H^* then belong to location-scale families based either on the "extreme value" distribution $[\exp(-e^{-x}),\ -\infty < x < \infty]$, the Fréchet distribution $[\exp(-x^{-\alpha}),\ x > 0,\ \alpha > 0]$, or the Weibull distribution $[\exp\{-(-x)^{\alpha}\},\ x < 0,\ \alpha > 0]$ (see Galambos 1987 for details). What may be less familiar, however, is the result of Pickands (1981) to the effect that the underlying copula associated with H^* is necessarily of the form (4). Given the longstanding concern of actuaries for the prediction of extreme events and their related costs, maximum extreme value copulas should be of considerable appeal in this area, where they would be expected to arise rather naturally. For a data-oriented presentation of univariate extreme value theory, with applications to insurance and other fields, see the recent book by Beirlant, Teugels and Vynckier (1996).

Of course, copulas in the class (4) can provide an appropriate model of dependence between variables, whether the marginals are of one of the above three types or not. The appropriateness of a family of extreme value copulas can be determined without knowledge of the marginals—as it should—using the procedure recently developed by Ghoudi, Khoudraji and Rivest (1998). Their test is based on the fact that if (X_1, X_2) is an observation from H^*, the distribution function K of $V = H^*(X_1, X_2)$ has the form $K(v) = v - (1 - \tau)v \log(v)$ for $0 \le v \le 1$, where

$$\tau = \int_0^1 \frac{t(1-t)}{A(t)}\, dA'(t)$$

is the population value of Kendall's tau. Then $\mu = 8E(V) - 9E(V^2) - 1 = 0$, and since the moments of V are easy to estimate, a test statistic, Z, can be defined that rejects model (4) if the sample version of μ, divided by a jackknife estimate of its standard deviation, is significantly different from zero, as compared to the standard normal law.

When applied to the data of Frees and Valdez, the test of Ghoudi, Khoudraji and Rivest (1998) yields a value of $z = 0.06$ that is clearly insufficient to reject the maximum extreme value copula model. Because the procedure is consistent for Archimedean alternatives, this provides additional evidence that Clayton's and Frank's families are inferior to Gumbel's in this case. This is not to say that the latter family is best within the class of extreme value copulas, however.

Since dependence functions are univariate, the search for the best copula model of the form (4) can proceed along similar lines as for Archimedean copulas. To choose between various parametric families of dependence functions A_γ that have been fitted to the data using the pseudo-likelihood approach of Genest, Ghoudi and Rivest (1995), one may plot the $A_{\hat\gamma}$'s against a nonparametric estimator A_n. Tawn (1988) suggests that the classical estimator of Pickands (1981) or the variant due to Deheuvels (1991) can be used for this purpose, but a much more efficient procedure has since been proposed by Capéraà, Fougères and Genest (1997).

Alternatively, the following proposition provides ways of embedding specific parametric families of dependence functions into larger systems, so that formal or informal selection procedures may be based as before on the corresponding maximized pseudo-log-likelihoods.

Proposition 3

Let A and B be two dependence functions.

(i) (*convex combination*) If $0 \le \lambda \le 1$, then $\lambda A + (1 - \lambda)B$ is a dependence function.

(ii) (*asymmetrization*) If $0 < \kappa, \lambda < 1$, and if $\bar u$ denotes $1 - u$ for arbitrary $0 \le u \le 1$, the following formula defines a dependence function:

$$E(t) = (\kappa t + \lambda \bar t)A\left(\frac{\kappa t}{\kappa t + \lambda \bar t}\right) + (\bar\kappa t + \bar\lambda \bar t)B\left(\frac{\bar\kappa t}{\bar\kappa t + \bar\lambda \bar t}\right),\ 0 \le t \le 1. \quad (5)$$

As with the previous propositions, the above list is not exhaustive. Rule (i), which is mentioned by Tawn (1988), is the special case of (ii) corresponding to $\kappa = \lambda$. Function (5) is the dependence function of the extreme value copula defined by

$$C_A(u^{1-\kappa}, v^{1-\lambda})C_B(u^{\kappa}, v^{\lambda}) \quad (6)$$

for all $0 \le u, v \le 1$. Rule (ii) is thus a consequence of the general asymmetrization process alluded to below the statement of Proposition 2.

To illustrate this final point, suppose that Gumbel's model, C_B, is enlarged via (6) using the independence copula $C_A(u, v) = uv$, generated by $A \equiv 1$. The resulting asymmetric model, already investigated at the end of the previous section, is then a three-parameter, maximum extreme value family of copulas, whose dependence functions are of the form

$$E(t) = \{\kappa t + \lambda(1 - t)\} + \{\kappa^\alpha t^\alpha + \lambda^\alpha(1 - t)^\alpha\}^{1/\alpha}, \quad 0 \leq t \leq 1.$$

Those who are familiar with the literature on multivariate extreme value theory will have recognized the class of dependence functions of what Tawn (1988) or Smith, Tawn and Yuen (1990) refer to as the asymmetric logistic model.

ACKNOWLEDGMENTS

This work was supported in part through research grants from the Natural Sciences and Engineering Research Council of Canada, from the Fonds pour la formation de chercheurs et l'aide à la recherche du Gouvernement du Québec, from the Fonds institutionnel de la recherche de l'Université du Québec à Trois-Rivières, and from the Fonds national de la recherche scientifique du Royaume de Belgique.

REFERENCES

ANDERSON, C.W., AND NADARAJAH, S. 1993. "Environmental Factors Affecting Reservoir Safety," in *Statistics for the Environment,* edited by V. Barnett and F. Turkman. New York: John Wiley.

BANDEEN-ROCHE, K.J., AND LIANG, K.-Y. 1996. "Modelling Failure-Time Associations in Data with Multiple Levels of Clustering," *Biometrika* 83:29–39.

BARBE, P., GENEST, C., GHOUDI, K., AND RÉMILLARD, B. 1996. "On Kendall's Process," *Journal of Multivariate Analysis* 58: 197–229.

BEIRLANT, J., TEUGELS, J.L., AND VYNCKIER, P. 1996. *Practical Analysis of Extreme Values.* Leuven: Leuven University Press.

CAPÉRAÀ, P., FOUGÈRES, A.-L., AND GENEST, C. 1997. "A Nonparametric Estimation Procedure for Bivariate Extreme Value Copulas," *Biometrika* 84:567–77.

CLAYTON, D.G. 1978. "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence," *Biometrika* 65:141–51.

DAY, R., BRYANT, J., AND LEFKOPOULOU, M. 1997. "Adaptation of Bivariate Frailty Models for Prediction, With Application to Biological Markers as Prognostic Indicators," *Biometrika* 84:45–56.

DEHEUVELS, P. 1991. "On the Limiting Behavior of the Pickands Estimator for Bivariate Extreme-Value Distributions," *Statistics and Probability Letters* 12:429–39.

FRANK, M.J. 1979. "On the Simultaneous Associativity of $F(x, y)$ and $x + y - F(x, y)$," *Aequationes Mathematicae* 19:194–226.

GALAMBOS, J. 1987. *The Asymptotic Theory of Extreme Order Statistics,* 2nd ed. Malabar, Fla.: Krieger.

GENEST, C., GHOUDI, K., AND RIVEST, L.-P. 1995. "A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions," *Biometrika* 82: 543–55.

GENEST, C., AND MACKAY, R.J. 1986. "Copules archimédiennes et familles de lois bidimensionnelles dont les marges sont données," *The Canadian Journal of Statistics* 14:145–59.

GENEST, C., AND RIVEST, L.-P. 1989. "A Characterization of Gumbel's Family of Extreme Value Distributions," *Statistics and Probability Letters* 8:207–211.

GENEST, C., AND RIVEST, L.-P. 1993. "Statistical Inference Procedures for Bivariate Archimedean Copulas," *Journal of the American Statistical Association* 88:1034–43.

GHOUDI, K., KHOUDRAJI, A., AND RIVEST, L.-P. 1998. "Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles," *The Canadian Journal of Statistics* 26:187–97.

GUMBEL, E. 1960. "Distributions des valeurs extrêmes en plusieurs dimensions," *Publications de l'Institut de statistique de l'Université de Paris* 9:171–73.

JOE, H. 1993. "Parametric Families of Multivariate Distributions with Given Margins," *Journal of Multivariate Analysis* 46: 262–82.

JOE, H., AND HU, T. 1996. "Multivariate Distributions from Mixtures of Max-Infinitely Divisible Distributions," *Journal of Multivariate Analysis* 57:240–65.

KHOUDRAJI, A. 1995. "Contributions à l'étude des copules et à la modélisation des valeurs extrêmes bivariées." Ph.D. thesis, Université Laval, Québec, Canada.

NELSEN, R.B. 1997. "Dependence and Order in Families of Archimedean Copulas," *Journal of Multivariate Analysis* 60:111–22.

OAKES, D. 1989. "Bivariate Survival Models Induced by Frailties," *Journal of the American Statistical Association* 84: 487–93.

OAKES, D. 1994. "Multivariate Survival Distributions," *Journal of Nonparametric Statistics* 3:343–54.

PICKANDS, J. 1981. "Multivariate Extreme Value Distributions," *Bulletin of the International Statistical Institute* 859–78.

SHIH, J.H., AND LOUIS, T.A. 1995. "Inferences on the Association Parameter in Copula Models for Bivariate Survival Data," *Biometrics* 51:1384–99.

SMITH, R.L., TAWN, J.A., AND YUEN, H.K. 1990. "Statistics of Multivariate Extremes," *International Statistical Review* 58: 47–58.

TAWN, J.A. 1988. "Bivariate Extreme Value Theory: Models and Estimation," *Biometrika* 75:397–415.

ZHENG, M., AND KLEIN, J.P. 1995. "Estimates of Marginal Survival for Dependent Competing Risks Based on an Assumed Copula," *Biometrika* 82:127–38.

DISCUSSIONS OF PAPERS ALREADY PUBLISHED

"UNDERSTANDING RELATIONSHIPS USING COPULAS," EDWARD FREES AND EMILIANO VALDEZ, JANUARY 1998

SHAUN S. WANG*

Recently the actuarial profession has witnessed a surge of interest in correlation models. This demand arises from many areas of actuarial practice, including modeling of some key financial market variables, multiple asset defaults, and catastrophe insurance losses. Thus, I was delighted to see the publication of the fine paper by Drs. Frees and Valdez; it provides an excellent review of state-of-the-art copula models.

In the last year I had the opportunity to undertake a research project on "Combining Correlated Risks" initiated by the Casualty Actuarial Society Committee on Theory of Risks. In doing that project, I investigated various models and methods for combining multiple correlated lines of business. Here I share some insights we gained into the correlation models and their relations to copula. In this discussion, I emphasize several points:

1. Drs. Frees and Valdez devote most of their discussion to the Archimedean family of copulas. From a practical point of view, I would like to advocate the use of the "normal copula." This non-Archimedean copula is especially powerful in modeling multiple (more than two) correlated variables. This is because only the normal copula allows an arbitrary correlation matrix yet still lends itself to efficient simulation techniques.
2. The copula formula was initially intended to be applied to cumulative distribution functions. Here I would like to show that copula formula can also be applied to probability-generating functions as well as characteristic functions. As a result of this innovative use of the copula formulas, we get some interesting multivariate distributions, which lead to efficient numerical techniques for calculating the aggregate loss distribution of correlated risks.
3. Although every correlation structure implicitly corresponds to a copula, the underlying copula may not be explicitly expressible. In fact, some very useful correlation models are best described in terms other than copula. I discuss some of these correlation models, especially with regard to modeling correlated claim-frequency variables.

1. THE NORMAL COPULA AND SIMULATION

In general, the modeling and combining of correlated risks are most straightforward if the correlated risks have a multivariate normal distribution. In this section, we use the multivariate normal distribution to construct the normal copula and then use that to generate multivariate distributions with arbitrary marginal distributions. The normal copula enjoys much flexibility in the selection of correlation parameters, and lends itself well to simple Monte Carlo simulation techniques.

Assume that $(Z_1, \cdots, Z_k)$ have a multivariate normal distribution with standard normal marginals $Z_j \sim N(0, 1)$ and a positive definite correlation matrix

$$\Sigma = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix},$$

where $\rho_{ij} = \rho_{ji}$ is the correlation coefficient between Z_i and Z_j. Then $(Z_1, \cdots, Z_k)$ have a joint p.d.f.:

$$f(z_1, \cdots, z_k) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left\{-\frac{1}{2} z' \, \Sigma^{-1} \, z\right\},$$

$$z = (z_1, \cdots, z_k). \quad (1.1)$$

From the correlation matrix Σ we can construct a lower triangular matrix

$$B = \begin{pmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{kk} \end{pmatrix}$$

such that $\Sigma = BB'$. In other words, the correlation matrix Σ equals the matrix product of B and its transpose B'. The elements of the matrix B can be calculated from the following Choleski's algorithm (Burden and Faires 1989, Sec. 6.6; Johnson 1987, Sec. 4.1):

*Shaun Wang, A.S.A., Ph.D., is an Associate Actuary at the SCOR Reinsurance Company, One Pierce Place, Itasca, Illinois 60143-4049, e-mail, scorus/itasca/swang%5512559@mcimail.com.

$$b_{ij} = \frac{\rho_{ij} - \sum_{s=1}^{j-1} b_{is}\, b_{js}}{\sqrt{1 - \sum_{s=1}^{j-1} b_{js}^2}}, \; 1 \le j \le i \le n, \qquad (1.2)$$

with the convention that $\Sigma_{s=1}^{0}(.) = 0$. Note that:

- For $i > j$, the denominator of Equation (1.2) equals b_{jj}.
- The elements of B should be calculated from top to bottom and from left to right.

The following simulation algorithm can be used to generate multivariate normal variables with a joint p.d.f. given by Equation (1.1) (Herzog 1986; Fishman 1996, pp. 223–24).

Step 1. Construct the lower triangular matrix $B = (b_{ij})$ by using Equation (1.2).

Step 2. Generate a column vector of independent standard normal variables $\mathbf{Y} = (Y_1, \cdots, Y_k)'$.

Step 3. Take the matrix product $\mathbf{Z} = B\mathbf{Y}$ of B and Equation $\mathbf{Y}$. Then $\mathbf{Z} = (Z_1, \cdots, Z_k)'$ has the required joint p.d.f. given by Equation (1.1).

Let $\Phi(.)$ represent the c.d.f. of the standard normal distribution

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt.$$

Then $\Phi(Z_1), \cdots, \Phi(Z_k)$ have a multivariate uniform distribution with Kendall's tau (Frees and Valdez 1998, p. 25)

$$\tau[\Phi(Z_i), \Phi(Z_j)] = \tau(Z_i, Z_j) = \frac{2}{\pi} \arcsin(\rho_{ij}),$$

and (Spearman) rank correlation coefficient

$$\text{RankCorr}(Z_i, Z_j) = \frac{6}{\pi} \arcsin\left(\frac{\rho_{ij}}{2}\right),$$

where $\arcsin(x)$ is an inverse trigonometric function such that $\sin(\arcsin(x)) = x$.

The following result can be easily verified but nevertheless is stated as a theorem due to its importance.

Theorem 1.1

Assume that $(Z_1, \cdots, Z_k)$ have a multivariate normal joint p.d.f. given by Equation (1.1), with correlation coefficient $\rho_{ij} = \rho(Z_i, Z_j)$. Let $H(z_1, \cdots, z_k)$ be their joint cumulative distribution function. Then

$$C(u_1, \cdots, u_k) = H[\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_k)]$$

defines a multivariate uniform c.d.f., the normal copula.

For any set of given marginal c.d.f.'s $F_1, \cdots, F_k$, the variables

$$X_1 = F_1^{-1}[\Phi(Z_1)], \cdots, X_k = F_k^{-1}[\Phi(Z_k)]$$

have a joint c.d.f.

$$F_{X_1,\cdots,X_k}(x_1, \cdots, x_k) = H\{\Phi^{-1}[F_1(x_1)], \cdots, \Phi^{-1}[F_k(x_k)]\}$$

with marginal c.d.f.'s $F_1, \cdots, F_k$. For the multivariate variables $(X_1, \cdots, X_k)$, we have Kendall's tau

$$\tau(X_i, X_j) = \tau(Z_i, Z_j) = \frac{2}{\pi} \arcsin(\rho_{ij})$$

and Spearman's rank correlation coefficient

$$\text{RankCorr}(X_i, X_j) = \text{RankCorr}(Z_i, Z_j) = \frac{6}{\pi} \arcsin\left(\frac{\rho_{ij}}{2}\right).$$

Although the normal copula does not have a simple analytical expression, it lends itself well to a very simple Monte Carlo simulation algorithm.

Suppose that we are given a set of correlated risks $(X_1, \cdots, X_k)$ with marginal c.d.f.'s $F_{X_1}, \cdots, F_{X_k}$ and Kendall's tau $\tau_{ij} = \tau(X_i, X_j)$, or rank correlation coefficient $\text{RankCorr}(X_i, X_j)$. If we assume that $(X_1, \cdot\cdot$ pd , $X_k)$ can be described by the normal copula in Theorem 1.1, then the following Monte Carlo simulation procedures can be used:

Step 1. Convert the given Kendall's tau or rank correlation coefficient to our usual measure of correlation for multivariate normal variables,

$$\rho_{ij} = \sin\left(\frac{\pi}{2}\,\tau_{ij}\right) = 2 \sin\left[\frac{\pi}{6}\,\text{RankCorr}(X_i, X_j)\right],$$

and construct the lower triangular matrix $B = (b_{ij})$ by Equation (1.2).

Step 2. Generate a column vector of independent standard normal variables $\mathbf{Y} = (Y_1, \cdots, Y_k)'$.

Step 3. Take the matrix product of B and $\mathbf{Y}$: $\mathbf{Z} = (Z_1, \cdots, Z_k)' = B\mathbf{Y}$. Set $u_i = \Phi(Z_i)$ for $i = 1, \cdots, k$.

Step 4. Set $X_i = F_{X_i}^{-1}(u_i)$ for $i = 1, \cdots, k$.

Theorem 1.1 and the associated simulation algorithm provide a powerful tool for generating correlated variables. The normal copula is very flexible because it allows any (symmetric, positive definite) matrix of rank correlation coefficients (or alternatively, Kendall's tau). The use of this algorithm implicitly assumes that the underlying multivariables can be described by a normal copula.

The normal copula is very practical because there are computer programs readily available for Monte Carlo simulation using a normal copula. For example, the Palisade Corporation product RISK, a Microsoft Excel add-in, implicitly utilizes the normal copula concept (Iman and Conver 1982).

2. AN INNOVATIVE USE OF THE COPULA FORMULA

The copula formula was initially intended to be applied to cumulative distribution functions. Here I demonstrate that, as a mathematical innovation, a copula formula can be applied to probability-generating functions or characteristic functions. With these new constructions, the resulting correlation structures lend themselves to efficient numerical algorithms for combining correlated risks (that is, calculating the aggregate loss distribution for the sum of the correlated risks).

2.1 Some Basics of Multivariables

In addition to joint cumulative distribution functions $F_{X_1,\cdots,X_k}(t_1, \cdots, t_k)$, other standard tools for multivariate random variables $(X_1, \cdots, X_k)$ include the joint p.g.f., joint m.g.f, and joint ch.f., which are defined as follows (see Johnson et al. 1997, pp. 2–12):

$$P_{X_1,\cdots,X_k}(t_1, \cdots, t_k) = \mathrm{E}[t_1^{X_1} \cdots t_k^{X_k}]$$

$$\phi_{X_1,\cdots,X_k}(t_1, \cdots, t_k) = \mathrm{E}[e^{\mathrm{i}(t_1X_1+\cdots+t_kX_k)}]$$

$$= P_{X_1,\cdots,X_k}(e^{\mathrm{i}t_1}, \cdots, e^{\mathrm{i}t_k}).$$

The joint p.g.f. $P_{X_1,\cdots,X_k}$ or the joint ch.f. $\phi_{X_1,\cdots,X_k}$ completely specifies the joint probability distribution. Equivalent results are obtained in terms of either a p.g.f. or a ch.f.

- The p.g.f. or ch.f. for the marginal (univariate) distribution F_{X_j} can be obtained by

$$P_{X_j}(t_j) = P_{X_1,\cdots,X_j,\cdots,X_k}(1, \cdots, 1, t_j, 1, \cdots, 1),$$

$$\phi_{X_j}(t_j) = \phi_{X_1,\cdots,X_j,\cdots,X_k}(0, \cdots, 0, t_j, 0, \cdots, 0).$$

- If the variables $X_1, \cdots, X_k$ are mutually independent, then

$$P_{X_1,\cdots,X_k}(t_1, \cdots, t_k) = \prod_{j=1}^{k} P_{X_j}(t_j).$$

- The covariances can be evaluated by $\mathrm{Cov}[X_i, X_j] = \mathrm{E}[X_iX_j] - \mathrm{E}[X_i]\mathrm{E}[X_j]$ with

$$\mathrm{E}[X_iX_j] = \frac{\partial^2}{\partial t_i\ \partial t_j} P_{X_1,\cdots,X_m}(1, \cdots, 1)$$

$$= -\frac{\partial^2}{\partial t_i\ \partial t_j} \phi_{X_1,\cdots,X_m}(0, \cdots, 0).$$

This can be seen from the expression

$$\frac{\partial^2}{\partial t_i \partial t_j} P_{X_1,\cdots,X_k}(t_1, \cdots, t_k) = \sum x_i x_j f_{x_1,\cdots,x_k}$$

$$(x_1, \cdots, x_k) t_1^{x_1} \cdots t_i^{x_i-1} \cdots t_j^{x_j-1} \cdots t_k^{x_k}.$$

The concepts of joint p.g.f. and joint ch.f. are very useful in combining correlated variables.

Theorem 2.2

For any k-correlated variables $X_1, \cdots, X_k$ with joint p.g.f. $P_{X_1,\cdots,X_k}$ and joint ch.f. $\phi_{X_1,\cdots,X_k}$, the sum $Z = X_1 + \cdots + X_k$ has a p.g.f. and a ch.f.

$$P_Z(t) = P_{X_1,\cdots,X_k}(t, \cdots, t), \quad \phi_Z(t) = \phi_{X_1,\cdots,X_k}(t, \cdots, t).$$

Proof $P_Z(t) = \mathrm{E}[t^{X_1+\cdots+X_k}] = \mathrm{E}[t^{X_1} \cdots t^{X_K}]$
$= P_{X_1,\cdots,X_k}(t, \cdots, t).$ ■

If we know the joint ch.f. of the k-correlated variables, it is straightforward to get the ch.f. for their sum

$$\phi_Z(t) = \phi_{X_1,\cdots,X_k}(t, \cdots, t).$$

Then the probability distribution of Z can be obtained by Fast Fourier Transform (FFT). The FFT is a mapping of n points to n points, which can be viewed as a discrete analog of the characteristic function. A relation in terms of ch.f. corresponds to a relation in terms of FFT. For more details of the FFT method, see Klugman, Panjer, and Willmot (1998).

Consider the aggregation of two correlated risk portfolios:

$$Z = (X_1 + \cdots + X_N) + (Y_1 + \cdots + T_K),$$

where N and K are correlated, the pair (N, K) is independent of the claim sizes X and Y, and the X_i's and Y_j's are mutually independent. We have

$$P_Z(t) = \mathrm{E}[t^Z] = \mathrm{E}[t^{(X_1+\cdots+X_N)+(Y_1+\cdots+Y_K)}]$$

$$= \mathrm{E}_{N,K}\mathrm{E}[t^{(X_1+\cdots+X_n)+(Y_1+\cdots+Y_m)} | N = n, K = m]$$

$$= \mathrm{E}_{N,K}[P_X(t)^N P_Y(t)^K]$$

$$= P_{N,K}[P_X(t), P_Y(t)].$$

In terms of ch.f. we have

$$\phi_Z(t) = P_{N,K}(\phi_X(t), \phi_Y(t)). \tag{2.3}$$

Having introduced some basic concepts, I now show how a copula formula can be applied to p.g.f.'s and ch.f.'s.

2.2 Multivariate Negative Binomial Distributions

Consider the following Cook-Johnson copula (Cook and Johnson 1981) with

$$C(u_1, u_2, \cdots, u_k) = \left\{\sum_{j=1}^{k} u_j^{-\alpha} - k + 1\right\}^{-1/\alpha}. \tag{2.4}$$

By applying the Cook-Johnson copula to the probability-generating functions, we have

$$P_{N_1,N_2,\cdots,N_k}(t_1, t_2, \cdots, t_k) = \left\{\sum_{j=1}^{k} P_{N_j}(t_j)^{-\alpha} - k + 1\right\}^{-1/\alpha}. \tag{2.5}$$

Wang (1998) shows that the joint p.g.f. in Equation (2.5) defines a multivariate negative binomial distribution if N_j has a negative binomial distribution with p.g.f.

$$P_{N_j}(t_j) = [1 - \beta_j(t_j - 1)]^{\alpha_j}$$

and provided that $0 \leq \alpha \leq \min[\alpha_j]$. In this multivariate negative binomial distribution, we have

$$Cov[N_i, N_j] = \alpha E[N_i]E[N_j].$$

2.3 Multivariate Long-Tailed Distributions

Consider the Morgenstein copula $C(u, v) = uv[1 + \alpha(1 - u)(1 - v)]$. By applying it to characteristic functions, we get

$$\phi_{X,Y}(s, t) = \phi_X(s)\phi_Y(t)\{1 + \alpha[1 - \phi_X(s)][1 - \phi_Y(t)]\}. \tag{2.6}$$

The bivariate ch.f. in Equation (2.6) can be used to construct bivariate lognormal distributions, bivariate Pareto distributions, or lognormal/Pareto pairs. This bivariate model lends itself to the FFT method of evaluating the sum

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)\{1 + \alpha[1 - \phi_X(t)][1 - \phi_Y(t)]\}.$$

3. SOME CORRELATION MODELS FOR CLAIM COUNTS

As noted by Drs. Frees and Valdez, parametric copula models may not be the best way of describing correlation models, especially for correlated frequency models.

In many situations, individual risks are correlated since they are subject to the same claim-generating mechanism or are influenced by changes in the common underlying economic or legal environment. For instance, in property insurance, risk portfolios in the same geographic location are correlated where individual claims are contingent on the occurrence and severity of a natural disaster (hurricane, tornado, earthquake, or severe weather condition). In liability insurance, new court rulings or social inflation may set new trends that affect the settlement of all liability claims for one line of business. Based on these considerations, I discuss two correlated frequency models: a common mixture model and a common shock model. None of these correlation models can be explicitly expressed in terms of copula.

3.1 A Common Mixture Model

Consider k discrete random variables $N_1, \cdots, N_k$. Assume that there exists a random parameter Θ such that

$$(N_j|\Theta = \theta) \sim \text{Poisson}(\theta\lambda_j), j = 1, \cdots, k,$$

where the variable Θ has a p.d.f. $\pi(\theta)$ and a moment-generating function M_Θ. For any given $\Theta = \theta$, the variables $(N_j|\theta)$ are independent and Poisson $(\lambda_j\theta)$ distributed with a conditional joint p.g.f.

$$P_{N_1,\cdots,N_k|\Theta}(t_1, \cdots, t_k|\theta) = \mathrm{E}[t_1^{N_1} \cdots t_k^{N_k}|\Theta = \theta] = e^{\theta[\lambda_1(t_1-1)+\cdots\lambda_k(t_k-1)]}.$$

However, unconditionally, $N_1, \cdots, N_k$ are correlated as they depend upon the same random parameter Θ. The unconditional joint p.g.f. for $N_1, \cdots, N_k$ is

$$\begin{aligned} P_{N_1,\cdots,N_k}(t_1, \cdots, t_k) &= \mathrm{E}_\Theta[\mathrm{E}(t_1^{N_1} \cdots t_k^{N_k}|\Theta)] \\ &= \int_0^\infty e^{\theta[\lambda_1(t_1-1)+\cdots+\lambda_k(t_k-1)]}\pi(\theta)d\theta \\ &= M_\Theta[\lambda_1(t_1 - 1) + \cdots + \lambda_k(t_k - 1)]. \end{aligned}$$

It has marginal p.g.f.'s $P_{N_j}(t_j) = M_\Theta(\lambda_j(t_j(- 1))$ with $\mathrm{E}[N_j] = \lambda_j \mathrm{E}[\Theta]$.

Note that

$$\begin{aligned}\mathrm{Cov}[N_i, N_j] &= \mathrm{E}_\Theta \mathrm{Cov}[N_i|\Theta, N_j|\Theta] \\ &\quad + \mathrm{Cov}[\mathrm{E}[N_i|\Theta], \mathrm{E}[N_j|\Theta]] \\ &= \mathrm{Cov}[\Theta\lambda_i, \Theta\lambda_j] = \lambda_i\lambda_j \,\mathrm{Var}[\Theta].\end{aligned}$$

In particular, if Θ has a gamma(α, 1) distribution with m.g.f. $M_\Theta(z) = (1 - z)^{-\alpha}$, then

$$\begin{aligned}&P_{N_1,\cdots,N_k}(t_1, \cdots, t_k) \\ &\quad = [1 - \lambda_1(t_1 - 1) - \cdots - \lambda_k(t_k - 1)]^{-\alpha} \qquad (3.1)\end{aligned}$$

defines a multivariate negative binomial with marginals NB(α, λ_j) and $\mathrm{Cov}(N_i, N_j) = \alpha\lambda_i\lambda_j$.

Consider combining k risk portfolios. Assume that the frequencies N_j, $j = 1, \cdots, k$, are correlated via a common Poisson-gamma mixture and have a joint p.g.f. given by Equation (3.1). If the severities X_j, $j = 1, \cdots, k$, are mutually independent and independent of the frequencies, there is a simple method of combining the aggregate loss distributions. Given

$$\lambda = \lambda_1 + \cdots + \lambda_k$$

and

$$P_X(t) = \frac{\lambda_1}{\lambda_1} P_{X_1}(t) + \cdots \frac{\lambda_1}{\lambda} P_{X_k}(t),$$

then

$$P_{N_1,\cdots,N_k}[P_{X_1}(t), \cdots, P_{X_k}(t)] = [1 - \lambda(P_X(t) - 1)]^{-\alpha.}$$

In other words, the total loss amount for the combined risk portfolios has a compound NB(α, λ) distribution with the severity distribution being a weighted average of individual severity distributions. In this case, dependency does not complicate the computation; in fact, it simplifies the calculation. It is simpler than combining independent compound negative binomial distributions.

Note that the common Poisson-gamma mixture model is a special case of the multivariate negative binomial family in Equation (2.5). In the Poisson-gamma mixture model, the k marginals NB(α, λ_j) are required to have the same parameter α. On the other hand, the multivariate negative binomial family in Equation (2.5) allows arbitrary negative binomial frequencies NB(α_j, λ_j).

3.2 A Common Shock Model

Consider a multivariate Poisson distribution defined by the following joint p.g.f.

$$\begin{aligned}&P_{N_1,N_2,N_3}(t_1, t_2, t_3) = \exp \\ &\left\{\sum_{i=1}^{3} \lambda_{ii}(t_i - 1) + \sum_{i<j} \lambda_{ij}(t_it_j - 1) + \lambda_{123}(t_1t_2t_3 - 1)\right\}. \\ &\qquad\qquad (3.2)\end{aligned}$$

Its marginal distributions are

$$N_j \sim \mathrm{Poisson}(\lambda_{123} + \sum_{i=1}^{3} \lambda_{ij}), j = 1, 2, 3,$$

and for $i \neq j$, $\mathrm{Cov}[N_i, N_j] = \lambda_{ij} + \lambda_{123}$.

We let

- $K_{ii} \sim \mathrm{Poisson}(\lambda_{ii})$, for $i = 1, 2, 3$
- $K_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$, for $1 \leq i < j \leq 3$
- $K_{ij} = K_{ji}$, for $1 \leq i, j \leq 3$
- $K_{123} \sim \mathrm{Poisson}(\lambda_{123})$
- $N_j = K_{1j} \oplus K_{2j} \oplus K_{3j} \oplus K_{123}$, for $j = 1, 2, 3$.

Then the so-constructed (N_1, N_2, N_3) have a joint p.g.f. given by (3.2). In this model, K_{123} represents the common shock among all three variables (N_1, N_2, N_3). In addition, for $i \neq j$, $K_{ij} = K_{ji}$ represents the extra common shock between N_i and N_j.

Note that we can simulate the correlated frequencies, (N_1, N_2, N_3), component by component.

Neither the common Poisson mixture or the common shock model can be explicitly expressed in terms of copula.

REFERENCES

BURDEN, R.L., AND FAIRES, J.D. 1989. *Numerical Analysis,* 4th ed. Boston: PWS-KENT Publishing Company.

COOK, R.D., AND JOHNSON, M.E. 1981. "A Family of Distributions for Modelling Non-elliptically Symmetric Multivariate Data," *Journal of the Royal Statistical Society* B 43:210–18.

FISHMAN, G.S. 1996. *Monte Carlo: Concepts, Algorithms, and Applications.* New York: Springer-Verlag.

FREES, E.W., AND VALDEZ, E.A. 1998. "Understanding Relationships Using Copulas," *NAAJ* 2, no. 1:1–25.

HERZOG, T.N. 1986. "An Introduction to Stochastic Simulation," Casualty Actuarial Society Study Note 4B.

IMAN, R.L., AND CONVER, W.J. 1982. "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables," *Communications in Statistics: Simulation Computation* 11, no. 3:311–34.

JOHNSON, M.E. 1987. *Multivariate Statistical Simulation.* New York: John Wiley and Sons, Inc.

JOHNSON, N., KOTZ, S., AND BALAKRISHNAN, N. 1997. *Discrete Multivariate Distribution.* New York: John Wiley and Sons, Inc.

KLUGMAN, S., PANJER, H.H., AND WILLMOT, G.E. 1998. *Loss Models: From Data to Decisions.* New York: John Wiley and Sons, Inc.