

Correlations and Copulas for Decision and Risk Analysis

by

Robert T. Clemen
Fuqua School of Business
Duke University
Durham, NC 27708
clemen@mail.duke.edu

and

Terence Reilly
Lundquist College of Business
University of Oregon
Eugene, OR 97403
reilly@euclid.uoregon.edu

March, 1997

Correlations and Copulas for Decision and Risk Analysis

Abstract

The construction of a probabilistic model is a key step in most decision and risk analyses. Typically this is done by defining a joint distribution in terms of marginal and conditional distributions for the model's random variables. We describe an alternative approach that uses a copula to construct joint distributions and pairwise correlations to incorporate dependence among the variables. The approach is designed specifically to permit the use of an expert's subjective judgments of marginal distributions and correlations. The copula that underlies the multivariate normal distribution provides the basis for modeling dependence, but arbitrary marginals are allowed. We discuss how correlations can be assessed using techniques that are familiar to decision analysts. The approach is demonstrated in the context of a simple example.

(Measures of Dependence, Spearman's τ , Copulas, Multivariate Normal Copula, Decision Analysis Process)

1. Introduction

One of the central steps in decision and risk analysis is the construction of a model that portrays the uncertainty inherent in the situation. For example, such uncertainty could relate to risks associated with hazardous chemicals, uncertainty due to economic variables, or the stochastic nature of a manufacturing or service process. The conventional approach to modeling uncertainty is to specify a joint distribution of the random variables as a product of marginal and conditional distributions.

A disadvantage with the typical marginal-and-conditional approach is that the number of probability assessments that must be made can grow exponentially with the number of variables. Analysts respond to this by searching diligently for conditional independence among variables in order to reduce the assessment burden. In this paper we discuss an alternative to the marginal-and-conditional modeling approach. In particular, we show how a joint distribution can be constructed using a copula, requiring only marginal distributions and measures of dependence among the random variables. Using the copula that underlies the multivariate normal distribution, a complete copula-based joint distribution can be constructed using assessed rank-order correlations and marginal distributions. The results are that the assessment burden may be reduced substantially and that the calculation of conditional distributions becomes straightforward. The further implication is that the search for conditional independence is no longer crucial.

The discussion in this paper relies fundamentally on the presumption that experts can reliably assess correlations. Although this appears to be a tall order, some practical correlation-assessment approaches are available. In Section 4 we will discuss several ways to assess rank-order correlations using techniques that are familiar to decision analysts.

We do not view the copula approach as either a panacea or a wholesale replacement for conventional probability modeling using marginals and conditionals. Rather, it is best thought of

as an additional useful tool for a decision or risk analyst. As with other approaches, this one has costs and limitations, and we discuss those below. For now, we note that our approach lends itself more readily to cases in which the uncertain variables are continuous rather than discrete. In addition, the use of correlation as a dependence measure limits the nature of the probabilistic relationships that can be modeled (although we will argue that such limitations may be reasonable).

In the next section, we discuss the basics of constructing a copula-based probability model, showing how this approach can be implemented using the multivariate-normal copula density. Section 3 provides a step-by-step demonstration of the procedure in the context of the Eagle Airlines example from Clemen (1996). Section 4 discusses issues in the assessment of correlations and makes recommendations for viable procedures. Section 5 addresses additional issues associated with copula-based modeling and concludes.

2. Copula-Based Probability Models

Our starting point is the same as it would be in a conventional decision analysis; that is, we assume that the analyst has identified those uncertain variables for which a probabilistic model is required. For example, the analyst might have gone through a deterministic sensitivity analysis (Howard & Matheson, 1983) or Reilly's (1996) dependent sensitivity analysis. With a set of random variables identified, the analyst can proceed with the assessments of marginal distributions and dependence measures necessary for constructing the copula-based joint distribution. Before we discuss specific copula models and assessment techniques, however, we review basic properties of copulas.

2.1. Copula basics. The essence of the copula approach is that the joint distribution of the random variables is defined as a function of the marginal distributions. To make this notion precise, we review two essential mathematical results. The first is:

Sklar's Theorem (1959): Given a joint distribution function $F(x_1, \dots, x_n)$ for random variables X_1, \dots, X_n with marginals $F_1(x_1), \dots, F_n(x_n)$, F can be written as a function of its marginals:

$$F(x_1, \dots, x_n) = C[F_1(x_1), \dots, F_n(x_n)],$$

where $C(u_1, \dots, u_n)$ is a joint distribution with uniform marginals.

Moreover, if each F_i is continuous, C is unique.

The function C is called a *copula*. Sklar's theorem is important because it tells us that any joint distribution can be written in copula form.

Note that Sklar's Theorem is written in terms of cumulative distribution functions (for which we will use the term *distribution function* to distinguish from density or mass functions in what follows). The second result we need is that, under the assumption that each F_i and C are differentiable, the joint density $f(x_1, \dots, x_n)$ can be written as

$$f(x_1, \dots, x_n) = f_1(x_1) \times \dots \times f_n(x_n) c[F_1(x_1), \dots, F_n(x_n)], \quad (1)$$

where $f_i(x_i)$ is the density corresponding to $F_i(x_i)$, and $c = \partial^n C / (\partial F_1 \dots \partial F_n)$ is called the *copula density*. Equation (1) states that, under appropriate conditions, the joint density can be written as a product of the marginal densities and the copula density. For example, if the X_i 's are independent, then $c = 1$ and $f(x_1, \dots, x_n) = f_1(x_1) \times \dots \times f_n(x_n)$, the familiar formula for n independent uniform random variables. In contrast, the conventional modeling approach decomposes the joint density as a product of marginals and conditionals. From the representation

in (1) it is clear that the copula density c encodes information about the dependence among the X_i 's. For this reason c is sometimes called a *dependence function*.

To demonstrate the use of a copula to specify a joint distribution with specified marginals, suppose the analyst wishes to construct a bivariate distribution $H(x, y)$ with marginals $F(x)$ and $G(y)$. Specifically, let $F(x)$ be a beta distribution with parameters $(\alpha=5, \beta=5)$, let $G(y)$ be a lognormal distribution with parameters $(\mu=0, \sigma^2=1)$, and let

$$C_d(u, v) = \frac{-1}{d} \ln \left(1 + \frac{(e^{-du} - 1)(e^{-dv} - 1)}{e^{-d} - 1} \right), d \neq 0.$$

Then $C_d[F(x), G(y)]$ is a bivariate distribution of the requisite form $H(x, y)$. The copula $C_d(u, v)$ is a member of Frank's family (Frank, 1979); the parameter d encodes the dependence between X and Y (Nelson, 1986). As $d \rightarrow 0$, $C_d[F(x), G(y)]$ approaches $F(x)G(y)$ implying independence. As $d \rightarrow \infty$, the correlation increases.

Figure 1 displays the joint density $h(x, y)$ for three different levels of dependence; in each case X and Y have the required beta and lognormal marginal densities. Figures 1a and 1b show a three-dimensional plot and a contour plot of the bivariate density when X and Y have Spearman rank correlation¹ of 0.25. Figures 1c and 1d illustrate via contour plots the effect of increasing the level of correlation to 0.50 and 0.90, respectively.

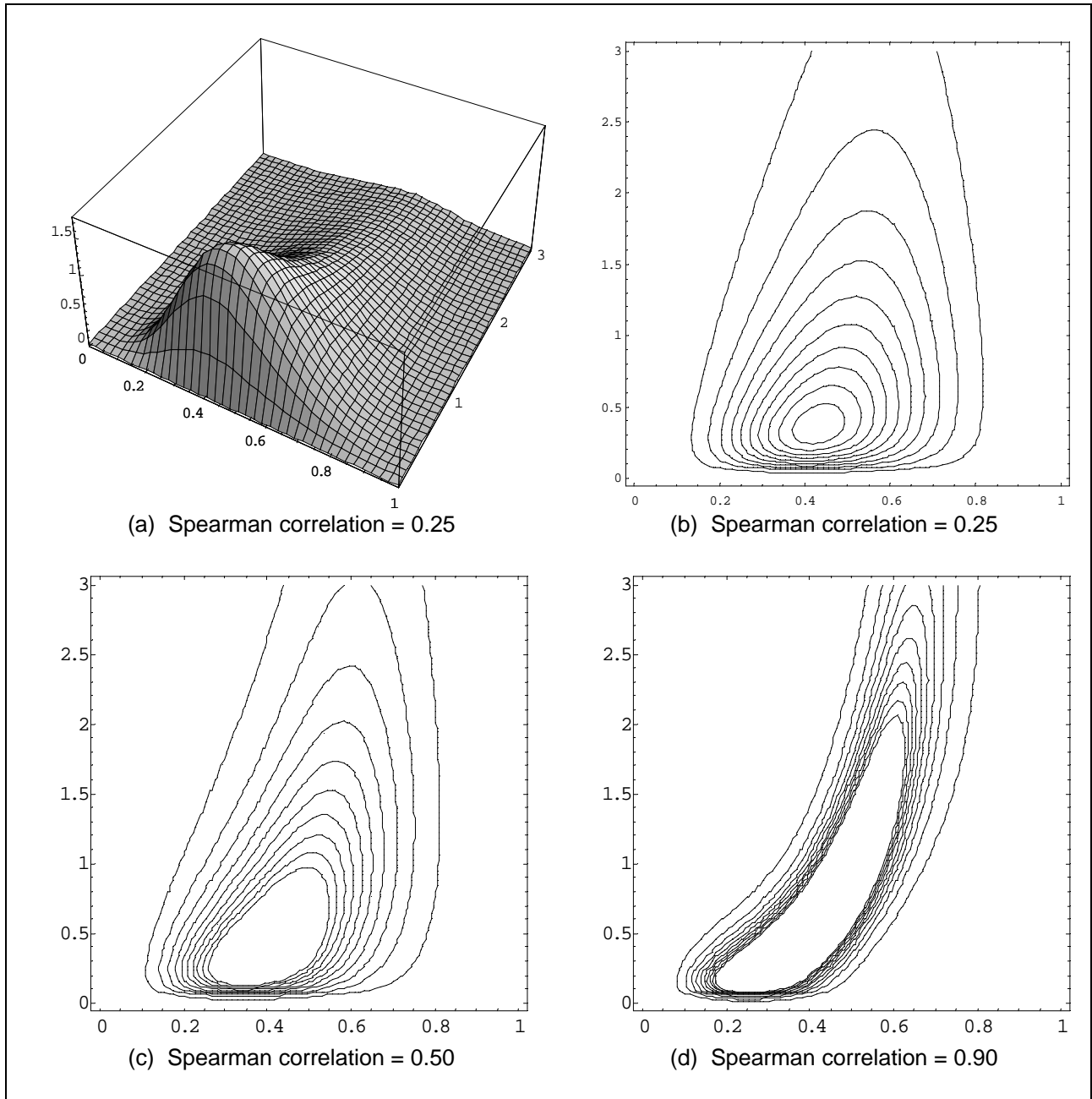


Figure 1. Bivariate density with beta (5, 5) and lognormal (0, 1) marginal densities.

Using a copula as a basis for constructing multivariate models is flexible because no restrictions are placed on the marginal distributions. Thus, we could have just as easily constructed a bivariate distribution with normal and gamma marginals or binomial and exponential marginals or any two subjectively assessed marginals. Such bivariate distributions are constructed by the substitution $u = F(x)$ and $v = G(y)$ as in the example above. Additional examples and background information on copulas can be found in Dall’Aglia, Kotz, and Salinetti (1991). Jouini and Clemen (1996) discuss the use of Archimedean copulas for aggregating expert judgments. Yi and Bier (1995) use copulas in an analysis of precursor events in a reliability model.

2.2. The multivariate normal copula. The fact that the copula is a function of the marginals is especially useful because it allows the “coupling” of the marginals into a joint distribution. Doing this requires two steps. First is modeling the marginal distributions in some way, which may require making a number of probability assessments and possibly fitting a member of a distribution family (normal, exponential, beta, etc.) to those assessments. Standard techniques from decision and risk analysis are available to accomplish these tasks (Morgan and Henrion, 1990; Clemen, 1996).

The second step is to create a copula that captures the dependence among the random variables. Although the relationships among the X_i ’s may be very complex, we believe that a reasonable approach is to begin with dependence relations of a monotonic nature. Many concepts of dependence or association are present in the literature. For our purposes, the most intuitive is the relatively strong notion of *regression dependence* (Lehmann, 1966): X_i is *positively (negatively) regression dependent* on X_j if $F(x_i | x_j)$ is non-increasing (non-decreasing) in x_j for all x_i . (Equivalently, Barlow and Proschan (1975) would say that X_i and X_j are *stochastically non-decreasing (non-increasing)*.) In our models we want to permit either positive or negative

dependence, and so we will use the terms *monotone regression dependence*, following Kelly and Krzysztofowicz (1996a).

In words, positive regression dependence means that increases in the conditioning variable imply that X_i is more likely to be greater, and *vice versa*; this is consistent with common intuition about regression-like relationships. If the expert believes that the dependence relationship among a pair of variables (or a suitable transformation thereof) satisfies the definition of monotone regression dependence, then we argue that a reasonable approach is to model that dependence with a rank-order correlation such as Spearman's rank correlation (r).² Otherwise, the analyst may default to the conventional representation using marginal and conditional distributions.

Rank-order correlations are useful because, unlike the Pearson product-moment correlation, they do not depend on the marginal distributions. For now, we will assume that the expert has assessed a matrix \mathbf{R}_S of Spearman rank-order correlations. Assessment of Spearman's r and related measures is discussed below.

Several copula families are available that can incorporate the relationships defined by matrix \mathbf{R}_S . One such family is the copula c_N that underlies the multivariate normal distribution. Like other copula families, the multivariate normal copula allows any marginal distribution for the X_i 's (beta, gamma, lognormal, etc.). It is called the normal copula because it encodes dependence in precisely the same way that the multivariate normal distribution does using only pairwise correlations among the variables, but it does so for variables with arbitrary marginals. Moreover, the normal copula permits the use of any positive-definite correlation matrix. (The class of Archimedean copulas, for example, is limited to intraclass correlation matrices; see Jouini and Clemen, 1996.) The flexibility and analytical tractability of the multivariate normal copula suggest that it is a promising way to represent dependence.

To understand the multivariate normal copula, begin by recalling that the multivariate normal distribution typically is parameterized in terms of Pearson product-moment correlations. Thus, for each element r_{ij} of \mathbf{R}_S , calculate the corresponding product-moment correlation r_{ij} for the multivariate normal as $r_{ij} = 2\sin(\pi r_{ij}/6)$ (Kruskal, 1958), and construct matrix \mathbf{R} with elements r_{ij} . Solving equation (1) for the copula density c_N in terms of the n -dimensional multivariate normal density $f^{(n)}(y_1, \dots, y_n | \mathbf{R})$ gives:

$$c_N[\Phi(y_1), \dots, \Phi(y_n) | \mathbf{R}] = f^{(n)}(y_1, \dots, y_n | \mathbf{R}) / [f(y_1) \times \dots \times f(y_n)], \quad (2)$$

where Φ and f denote the univariate standard normal distribution and density, respectively.

Substitution of the expressions for the normal densities and algebraic manipulation lead to:

$$c_N[\Phi(y_1), \dots, \Phi(y_n) | \mathbf{R}] = \exp\{-\mathbf{y}^T (\mathbf{R}^{-1} - \mathbf{I}) \mathbf{y} / 2\} / |\mathbf{R}|^{1/2}, \quad (3)$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$, and \mathbf{I} is the $n \times n$ identity matrix.

We are now able to construct a multivariate density using c_N as a dependence function with arbitrary marginals $F_1(x_1), \dots, F_n(x_n)$. Using the normal inverse transformation Φ^{-1} , define $Y_i = \Phi^{-1}[F_i(X_i)]$ for $i=1, \dots, n$, and substitute these into (3) and (1) to obtain the desired joint density:

$$\begin{aligned} f(x_1, \dots, x_n | \mathbf{R}) &= f_1(x_1) \times \dots \times f_n(x_n) \times \exp\{-\mathbf{y}^T (\mathbf{R}^{-1} - \mathbf{I}) \mathbf{y} / 2\} / |\mathbf{R}|^{1/2} \\ &= f_1(x_1) \times \dots \times f_n(x_n) \times \\ &\quad \exp\{-\left(\Phi^{-1}[F_1(x_1)], \dots, \Phi^{-1}[F_n(x_n)]\right) (\mathbf{R}^{-1} - \mathbf{I}) \left(\Phi^{-1}[F_1(x_1)], \dots, \Phi^{-1}[F_n(x_n)]\right)^T / 2\} / |\mathbf{R}|^{1/2} \end{aligned} \quad (4)$$

This joint density has the specified marginals and, because Spearman's r is invariant under monotone 1-1 transformations of the original variables, the X_i 's have the assessed rank-order correlations \mathbf{R}_S . Calculating the density for specific values x_1, \dots, x_n is relatively easy, requiring n inversions of the univariate standard normal distribution.

Conditional densities are also easily calculated using the multivariate normal copula model.

Let \mathbf{R} and \mathbf{y} be partitioned as follows:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{n-1} & \mathbf{r} \\ \mathbf{r}^T & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = (\mathbf{y}_{n-1}, y_n),$$

where $\mathbf{y}_{n-1} = (y_1, \dots, y_{n-1})^T$, \mathbf{R}_{n-1} is the $(n-1) \times (n-1)$ correlation matrix for (Y_1, \dots, Y_{n-1}) , and $\mathbf{r} = (r_{1n}, \dots, r_{(n-1)n})^T$. From (3), the definition of conditional probability, and the formula for the standard multivariate normal density, we have

$$\begin{aligned} & f(x_n | x_1, \dots, x_{n-1}, \mathbf{R}_S) \\ &= f_n(x_n) \times \frac{f^{(n)}(\Phi^{-1}[F_1(x_1)], \dots, \Phi^{-1}[F_n(x_n)] | \mathbf{R})}{f(\Phi^{-1}[F_n(x_n)]) \times f^{(n-1)}(\Phi^{-1}[F_1(x_1)], \dots, \Phi^{-1}[F_{n-1}(x_{n-1})] | \mathbf{R}_{n-1})}, \end{aligned}$$

which, upon substituting in the expressions for the normal densities and reducing, becomes

$$\begin{aligned} & f(x_n | x_1, \dots, x_n, \mathbf{R}) \\ &= f_n(x_n) \exp \left\{ -0.5 \left[\frac{(\Phi^{-1}[F_n(x_n)] - \mathbf{r}^T \mathbf{R}_{n-1}^{-1} \mathbf{y}_{n-1})^2}{(1 - \mathbf{r}^T \mathbf{R}_{n-1}^{-1} \mathbf{r})} - (\Phi^{-1}[F_n(x_n)])^2 \right] \right\} (1 - \mathbf{r}^T \mathbf{R}_{n-1}^{-1} \mathbf{r})^{-1/2}. \quad (5) \end{aligned}$$

Although the formulas above appear complex, the resulting expressions in (4) and (5) are easy to use in computations. For example, the conditional density in (5) is the product of a marginal density, an exponential term, and a scaling constant.

With the joint density specified, expected values, expected utilities, and risk profiles can be calculated directly from the copula model. Value-of-information analysis is possible and will typically require the calculation of conditional distributions from the copula in a manner similar to (5) above. As usual, modeling and analysis may iterate until clarity of action is obtained. The example in Section 3 demonstrates both discrete-approximation and simulation approaches for performing some of the calculations that might be required in an analysis.

3. An Example: Eagle Airlines

3.1. Initial model specification and deterministic sensitivity analysis. Clemen (1996) describes the hypothetical decision faced by Dick Carothers, owner of the fledgling Eagle Airlines. Carothers is considering purchasing a used aircraft. His decision criterion is whether the airplane will generate more profit than a money-market alternative investment. Reilly (1996) modifies the model slightly in his sensitivity-analysis example, and we will use Reilly's model here. The influence diagram in Figure 2 portrays the deterministic model, and Table 1 lists Carother's low, base, and high values for the input variables. For illustrative purposes, we take these to be the assessed 10th, 50th, and 90th percentiles of the respective probability distributions.

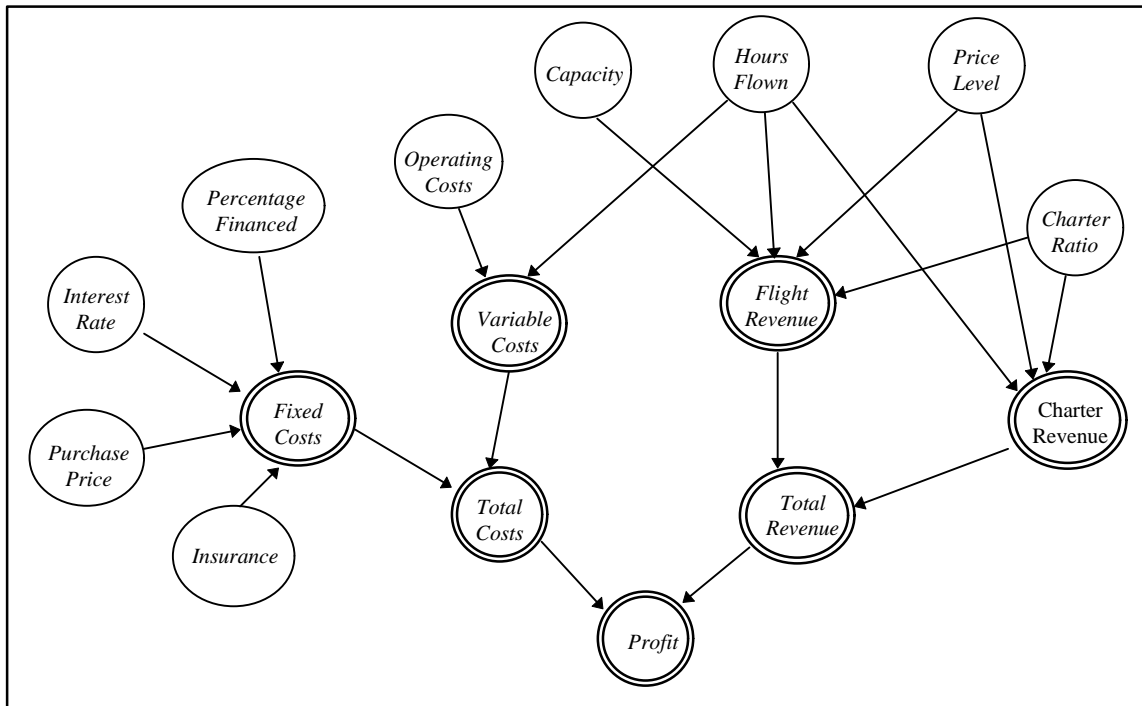


Figure 2. The initial influence diagram for Eagle Airlines.

Variable (X)	Fractile:	Low	Base	High
Charter Ratio		45%	50%	70%
Capacity		40%	50%	60%
Price Level		\$95	\$100	\$108
Hours Flown		500	800	1000
Operating Cost per Hour		\$230	\$245	\$260

<i>Percentage Financed</i>	30%	40%	50%
<i>Interest Rate</i>	10.50%	11.50%	13.00%
<i>Insurance</i>	\$18,000	\$20,000	\$25,000
<i>Purchase Price</i>	\$85,000	\$87,500	\$90,000

Table 1. Input variables and assessed fractiles for use in Eagle Airlines sensitivity analysis.

The first step is to conduct a sensitivity analysis to identify the critical variables. As indicated above, this can be accomplished either with the standard one-way sensitivity analysis (Howard and Matheson, 1983; Clemen, 1996) or dependent sensitivity analysis as described by Reilly (1996). As Reilly demonstrates, in this case the two sensitivity-analysis approaches produce slightly different insights but eventually identify the same set of critical variables. Throughout we will denote *Price*, *Hours Flown*, *Capacity*, and *Operating Cost* by *P*, *H*, *C*, and *O*, respectively. Information on these variables, including example correlations, are shown in Table 2. These correlations would have been assessed as part of the dependent sensitivity analysis, or in a subsequent assessment step if the analyst has followed the standard one-way sensitivity analysis.

Variable (X)	Fractile:	Correlations					
		Low	Base	High	<i>Price Level</i>	<i>Hours Flown</i>	<i>Capacity</i>
<i>Price Level (P)</i>		\$95	\$100	\$108			
<i>Hours Flown (H)</i>		500	800	1000	-0.50		
<i>Capacity (C)</i>		40%	50%	60%	-0.25	0.50	
<i>Operating Cost per Hour (O)</i>		\$230	\$245	\$260	0	0	0.25

Table 2. Fractiles and Spearman correlations for four critical variables in Eagle Airlines

3.2. A copula-based joint density. To create a copula-based joint density, two steps are required. First is to model the marginal densities for the variables. To demonstrate, we fit marginal densities to the four variables as indicated in Table 3 so that the modeled densities have approximately the same 0.10, 0.50, and 0.90 fractiles as shown in Table 2. Denote the marginal density and cumulative distribution for *P* as $f_b(p)$ and $F_b(p)$, respectively, and similarly for *H* and

C .³ Likewise, let $f_N(o)$ and $F_N(o)$ denote the marginal density and cumulative distribution for the normally distributed O .

Variable	Distribution	Parameters	Range
Price Level (P)	Scaled beta	$a = 9, b = 15$	[\$81.94, \$133.96]
Hours Flown (H)	Scaled beta	$a = 4, b = 2$	[66.91, 1135.26]
Capacity (C)	Beta	$a = 20, b = 20$	[0, 1]
Operating Cost (O)	Normal	$m = 245, s = 11.72$	$(-\infty, +\infty)$

Table 3. Marginal distributions for Eagle Airlines probability model

The second step in creating the the joint density is to specify the copula c_N . This is a matter of using the assessed Spearman correlations to calculate \mathbf{R} and substituting \mathbf{R} and the marginal densities and distributions into (4). For Eagle airlines, we substituted the marginal densities from Table 3, the corresponding distributions (F_i), and \mathbf{R} into (4). Let $Y_P = \Phi^{-1}[F_b(P)]$, and similarly for H, C , and O , and let $\mathbf{y} = (y_P, y_H, y_C, y_O)$. We have:

$$f(p, h, c, o) = f_b(p)f_b(h)f_b(c)f_N(o) \exp\{-\mathbf{y}'(\mathbf{R}^{-1} - \mathbf{I})\mathbf{y} / 2\} / 0.486, \quad (7)$$

where

$$\mathbf{R}^{-1} - \mathbf{I} = \begin{bmatrix} 0.366 & 0.715 & -0.014 & 0.004 \\ 0.715 & 0.777 & -0.787 & 0.205 \\ -0.014 & -0.787 & 0.506 & -0.393 \\ 0.004 & 0.205 & -0.393 & 0.103 \end{bmatrix}.$$

Equation (7) represents the full multivariate joint density and can now be used in the analysis.

No convention currently exists for representing a copula-based joint distribution in an influence diagram. Any convention adopted must accommodate the presence of predecessors and successors that are external to the copula model. At the same time, directed arcs among the variables that are related by the copula are inappropriate. We suggest placing the copula-related variables in proximity to each other and within a shaded region as exemplified in Figure 3. This

convention suggests that each variable in the joint distribution will require assessment of its marginal distribution, and that measures of dependence among the variables will be necessary to specify the copula.

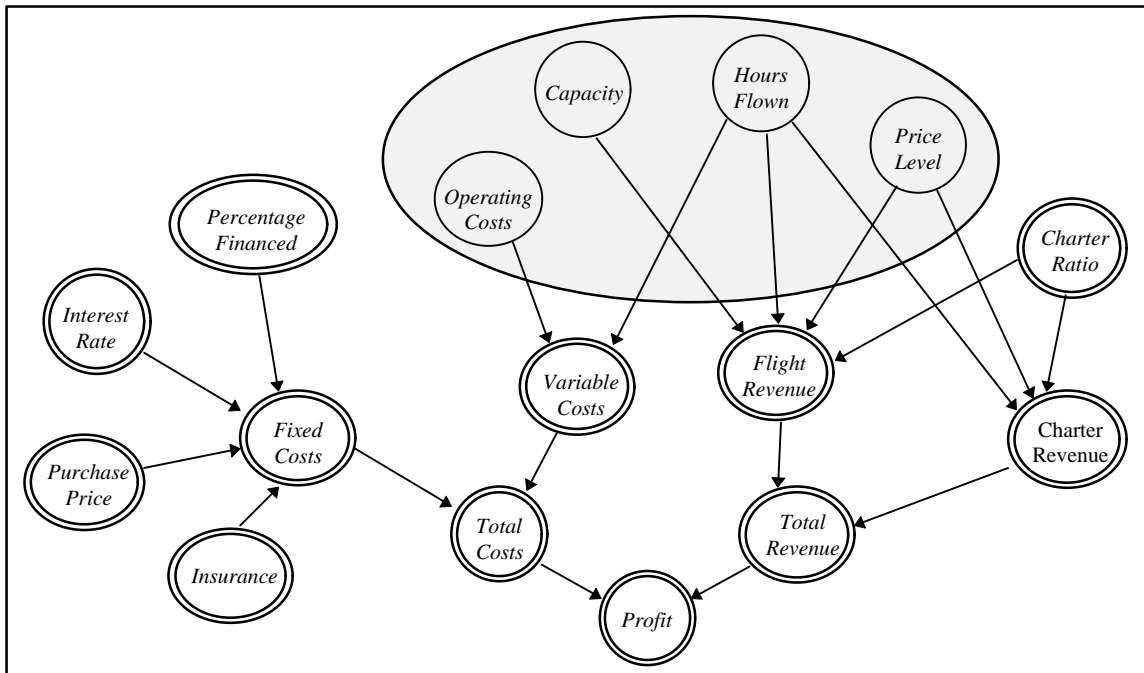


Figure 3. Representing a copula-based probability model in an influence diagram.

3.3. A Discrete Approximation. Analysis of the four-dimensional continuous density in (7) require numerical integration or Monte Carlo simulation. In the spirit of current practice in decision analysis, we show in this section how a discrete approximation can be created. The approach we take is to calculate conditional densities using (5) and use these to create an event tree based on the extended Pearson-Tukey discrete approximation procedure (Keefer and Bodily, 1983).

Using (5), it is straightforward to derive the conditional densities $f(h | p)$, $f(c | h, p)$, and $f(o | c, h, p)$:

$$f(h|p) = f_b(h) \exp\{-0.5[(\Phi^{-1}[F_b(h)] + 0.518\Phi^{-1}[F_b(p)])^2 / 0.732 - (\Phi^{-1}[F_b(h)])^2] \} / 0.732^{1/2}$$

$$f(c|p,h) = f_b(c) \exp\{-0.5[(\Phi^{-1}[F_b(c)] - 0.009\Phi^{-1}[F_b(p)] - 0.523\Phi^{-1}[F_b(h)])^2 / 0.732 - (\Phi^{-1}[F_b(c)])^2] \} / 0.732^{1/2}$$

$$f(o|p,h,c) = f_N(o) \exp\{-0.5[(\Phi^{-1}[F_N(o)] + 0.003\Phi^{-1}[F_b(p)] + 0.186\Phi^{-1}[F_b(h)] - 0.357\Phi^{-1}[F_b(c)])^2 / 0.907 - (\Phi^{-1}[F_N(o)])^2] \} / 0.907^{1/2}$$

These expressions are no more than realizations of Equation (5) in which we have substituted in the correlation values from **R** and the marginal densities from Table 3.

As indicated, each conditional density is approximated using the extended Pearson-Tukey method. Figure 4 shows a portion of the resulting event tree. The branch values displayed are the 0.05, 0.50, and 0.95 fractiles of each marginal or conditional density, and the probabilities on the top, middle, and bottom in each set of three branches are 0.185, 0.63, and 0.185, respectively, *per* the extended Pearson-Tukey formula. The complete event tree is available on request from the authors.

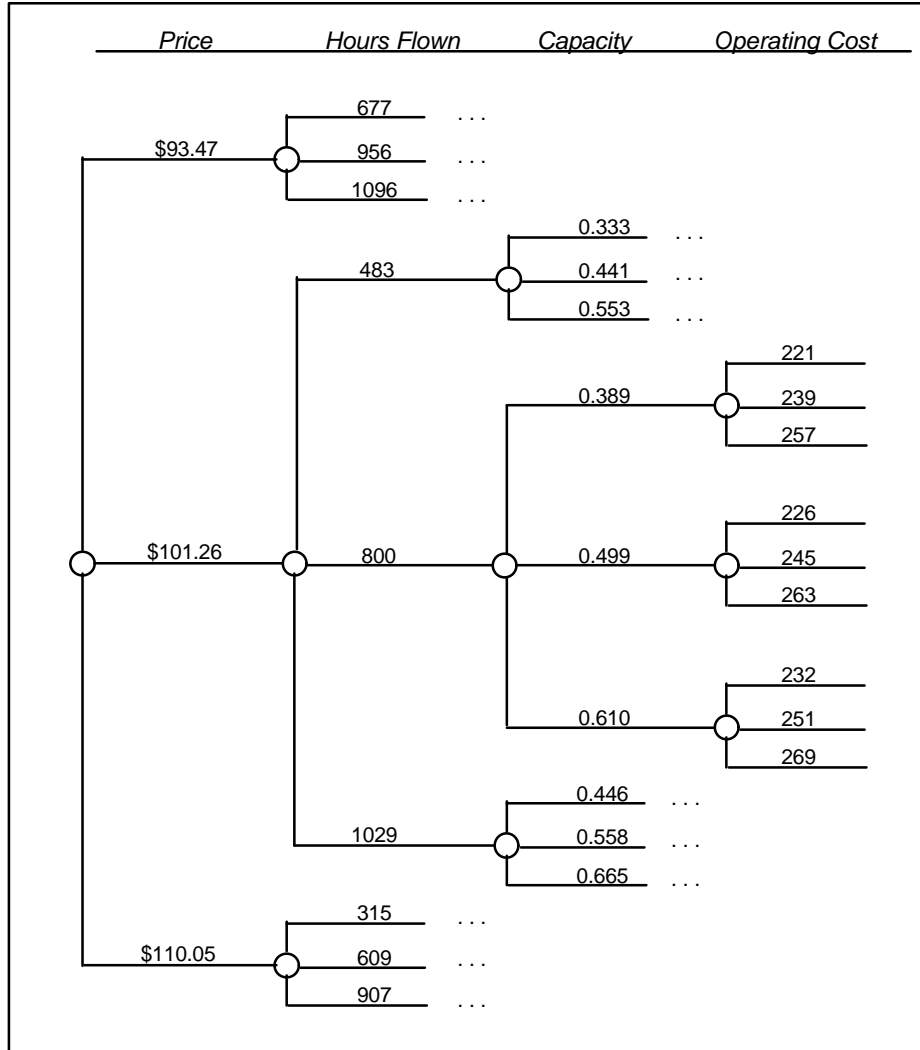


Figure 4. Event tree from multivariate normal copula.

The bivariate relationships among the variables reveal themselves readily in Figure 4. For example, as *Price* increases, *Hours Flown* tends to decrease, reflecting the negative correlation between these two variables. As *Hours Flown* increases, its positive correlation with *Capacity* leads to an increase in the latter. The same is true when we examine changes in the distribution of *Operating Cost* given *Capacity*.

3.4. Analysis. We can analyze the Eagle Airlines problem by using the multivariate normal copula model (7) directly in a Monte Carlo simulation. A vector variate (p, h, c, o) is generated by

first generating a vector (y_P, y_H, y_C, y_O) from a multivariate-normal process with correlation matrix \mathbf{R} . These variables are marginally standard normal, but considered jointly they have the rank correlations as specified in the model. The standard normal distribution function $\Phi(y_i)$ is calculated for each of the four y variables. Finally, we use the inverse marginal distribution functions to calculate $(p, h, c, o) = (F_P^{-1}[\Phi(y_P)], F_H^{-1}[\Phi(y_H)], F_C^{-1}[\Phi(y_C)], F_O^{-1}[\Phi(y_O)])$. This vector of variables comes from a process that has the specified marginal distributions as well as the required rank correlations.

It is instructive to compare the process described above to that of Iman and Conover (1982). The copula-based procedure generates variates from a process that is known to have the specified population parameters. Iman and Conover's procedure uses a complex set of steps designed to create a *sample* of observations for which the calculated *sample* rank correlation is close to the specified correlation. It is not known at this time what the population characteristics are for Iman and Conover's procedure. It would appear that the copula-based approach would be worth considering for Monte Carlo simulation programs like @RISK and Crystal Ball that permit users to specify correlated variables with arbitrary marginals.

We used the simulation procedure described above to calculate the expected profit, standard deviation, and risk profile for Eagle Airlines. For comparison purposes, we also ran the simulation under the assumption that the variables are independent. Both simulations were run for 10,000 trials. We also calculated the expected profit, standard deviation, and risk profile using the discrete approximation from Section 3.3.

The results are displayed in Table 4 and Figure 5. If we take the copula-based simulation as the benchmark, several observations can be made:

- Incorporating dependence can be important. The simulation that treated the variables as independent resulted in a risk profile with more variability.
- The discrete approximation closely mimics the benchmark in terms of both the expected profit and variability.
- The risk profile for the discrete approximation does a reasonably good job of replicating the copula simulation risk profile. Especially in the tails, the discrete approximation is nearly identical to the benchmark, a useful property because so many inferential procedures are sensitive to tail behavior.

4. Assessing Correlations

Application of the multivariate normal copula as described above requires the expert to assess correlations. Although useful methods for assessing probabilities are well known, analysts typically do not try to assess moments of distributions (Morgan and Henrion, 1990). The assessment of cross moments would appear to be an even more difficult problem. Although such judgments are inherently difficult, three general approaches are available to help an expert think carefully about relationships among random variables.

4.1. Statistical approaches. First are techniques that rely on an expert's familiarity with statistical concepts related to correlation. For example, an expert might make a judgment regarding the "percentage of variance explained" (r^2) that would result from regressing one variable on another. Another possibility is to have the expert view several scatterplots representing different levels of correlation and select a level of correlation that is consistent with the expert's belief about the strength of the relationship between the variables. This approach is

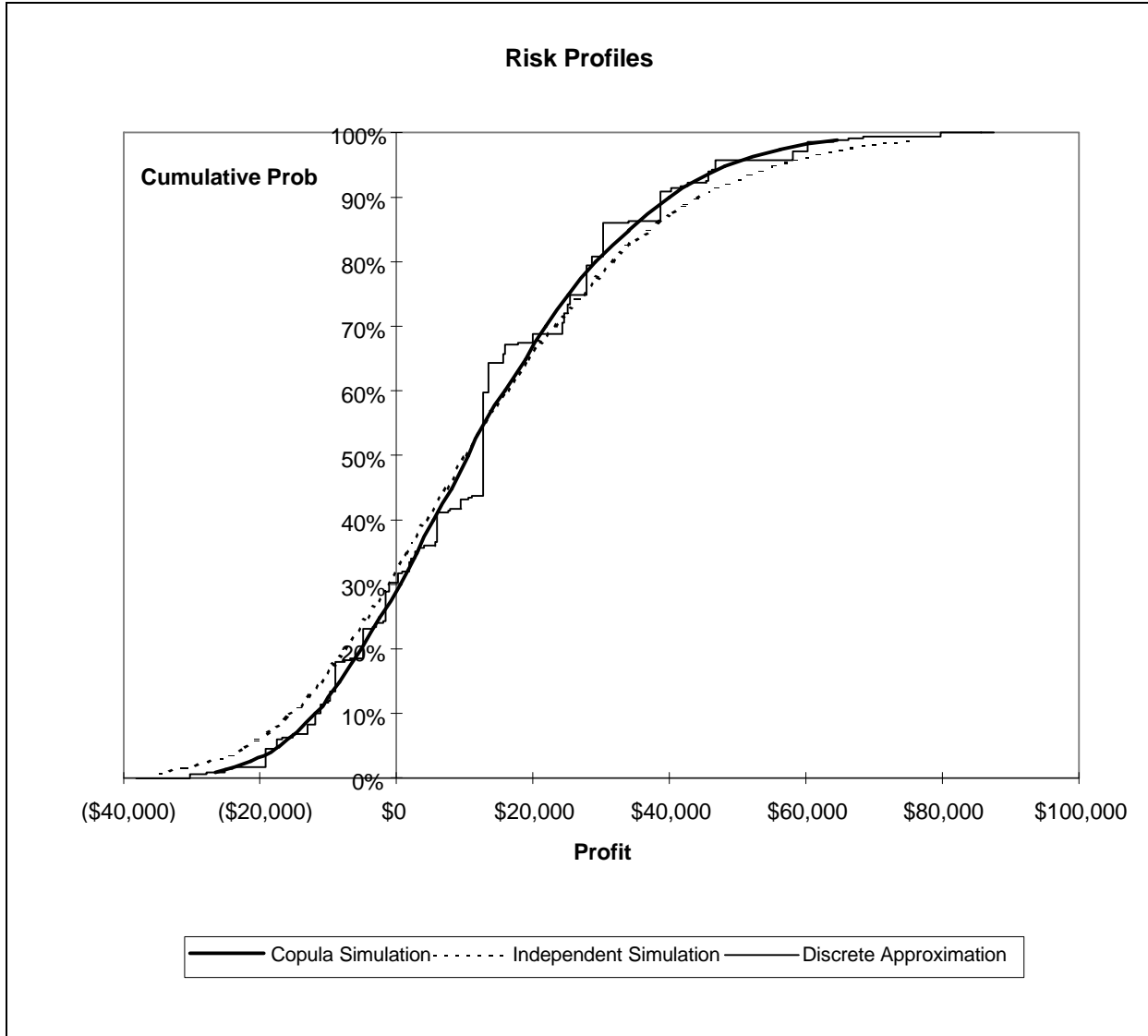


Figure 5. Risk profiles for Eagle Airlines from three models.

	Expected Value	Standard Deviation
Copula simulation (10,000 trials)	\$12,417	\$20,206
Independent simulation (10,000 trials)	\$12,426	\$23,628
Discrete approximation	\$12,606	\$20,281

Table 4. Expected profit and standard deviation for Eagle Airlines calculated from three models.

currently used informally in *Crystal Ball* and @RISK, the popular risk-analysis add-ins for Microsoft® Excel™.

The strength of these approaches is that experts who have substantial training in statistical data analysis may be able to think of bivariate relationships easily in these terms. For example, one of the authors (Clemen) has found in practice that experts sometimes spontaneously suggest the r^2 notion when thinking about bivariate relationships and are willing to give an assessment in terms of percentage of variance of one variable that is explained by the second. Gokhale and Press (1982) showed that individuals with statistical training were capable of viewing a scatterplot of bivariate data and making a reasonably accurate assessment of the sample correlation of those data. In either case, the accuracy of the expert's judgment depends on the depth of his or her familiarity with the notions of correlation or regression and an ability to match his or her belief regarding the strength of a bivariate relationship with an appropriate scatterplot or r^2 . Although the procedures have some intuitive appeal and "look easy," their psychometric validity has not been established.

4.2. Probability of concordance. The second general approach to correlation assessment is more familiar to decision analysts: Assess conditional or joint probabilities and relate those to the required measure of dependence. For example, Gokhale and Press had their subjects make assessments of the probability of concordance. For a bivariate population (X, Y) , we can define the concordance probability by considering two independent draws (x_1, y_1) and (x_2, y_2) . Concordance probability P_C is defined as:

$$\begin{aligned} P_C &= P[(x_1 \leq x_2 \text{ and } y_1 \leq y_2) \text{ or } (x_1 > x_2 \text{ and } y_1 > y_2)] \\ &= P(x_1 \leq x_2 \mid y_1 \leq y_2). \end{aligned}$$

P_C can be related to Kendall's τ :

$$t = 2 P_C - 1.$$

Kendall's t is a measure of dependence that, like Spearman's r , is invariant with respect to the marginal distributions. Thus, if an expert is able to assess a matrix of pairwise t_{ij} 's for variables (X_i, X_j) , this matrix can be used as a basis for specifying an appropriate copula. For example, the matrix \mathbf{R} for the multivariate normal copula can be defined on the basis of Kendall's t instead of Spearman's r , using the relationship $r_{ij} = \sin(\rho t/2)$. Jouini and Clemen (1996) describe the use of Kendall's t in the context of Archimedean copulas.

The primary difficulty with using the probability of concordance and t is the nature of the joint or conditional event for which the probability must be assessed. As long as there is a natural interpretation of the event in terms of frequency, this is a minor problem. For example, judging the relationship between height and weight in a population by assessing P_C would require the expert to answer a question like the following:

“Two individuals (labeled A and B) are chosen randomly from a population of adult males. Given that A weighs less than B, what is the probability that A is also shorter than B?”

Consider, however, what the assessment question would look like if we were interested in the relationship between the size (S) of a population of organisms (say, an endangered species of tree frogs) and a temperature index T for the tree frogs' environment. In such a case, there may be only one population of such frogs and no obvious frequency interpretation to relate S and T . The expert would have to answer a difficult question like the following:

“Imagine a hypothetical situation in which there are two separate and independent populations of tree frogs, each one with its own population size and temperature measurement. Call these two

situations A and B, with corresponding population size and temperature pairs (s_A, t_A) and (s_B, t_B) . Given that t_A is less than t_B in this hypothetical situation, what is the probability that s_A is also less than s_B ?”

Direct assessment of probability P_C may be appealing to the purist. As long as the assessment can be given a frequency or cross-sectional interpretation, the task seems quite reasonable. Many such situations come to mind, such as judgments of the relationships among variables in an epidemiological study of cancer risks, say, or judging relationships among the rates of returns of different projects or portfolios. Moreover, the frequency framing suggests that the expert would be less susceptible to cognitive biases (Gigerenzer, 1991; Gigerenzer, Hoffrage, & Kleinbolting, 1991). Many applications involve one-time events, such as the extinction of a species or global warming, and these situations may not have natural frequency interpretations. Thus, the practical value of using t and P_C may be limited.

4.3. Conditional fractile estimates. The third approach involves making conditional estimates and using these to derive Spearman’s r . Suppose we have random variables X and Y with corresponding distribution functions $F(x)$ and $G(y)$. The standard nonparametric regression representation between the fractiles of F and G is

$$E[F(X) | y] = r_{XY} [G(y) - 0.5] + 0.5, \quad (6)$$

where r_{XY} is the Spearman rank-order correlation between X and Y . Equation (6) suggests that the expert be asked questions like the following:

“Suppose an individual is randomly chosen from a population of adult males. Given that his height is found to be 180 cm, the 70th

percentile of the distribution of heights for this population, what would you estimate for the percentile at which his weight falls?"

This question asks the expert to make a conditional estimate of the percentile at which the individual's weight would fall. If the marginal distributions have been assessed, as would have been done in a typical analysis, then the percentile estimate can be related directly to the marginal distribution; for example, the expert might respond that the conditional estimate of the weight would be the 60th percentile of the marginal distribution of weights, corresponding to 85 kg. Given $G(180) = 0.70$ and $E[F(X) | 180] = 0.60 = F(85)$, the analyst can solve for r_{XY} . Furthermore, it would be natural to have the expert make several conditional estimates and use a least-squares approach to estimate r_{XY} .

Note that it is tempting to ask the expert to assess $E(X | y)$ and then to calculate $F[E(X | y)]$. That is, ask the expert for an estimate of an individual's weight given height = 180 cm, and calculate the percentile of the response from the marginal distribution for weight. Because of the function $F(X)$ is nonlinear, however, $F[E(X | y)] \neq E[F(X) | y]$, so strictly speaking this approach would be incorrect. We speculate, however, that experts will more naturally think in terms of the original variables, and that assessing several values of $E(X | y)$, using $F[E(X | y)]$ as an approximation of $E[F(X) | y]$, and using a least-squares approach to estimate r_{XY} will result in minimal distortion of the expert's judgment.

The conditional fractile approach described above is closely related to a technique called *predictive assessment* originally developed by Bayesian statisticians for assessing a prior distribution for the parameters in a linear model (Winkler, Smith, and Kulkarni, 1978; Kadane et al., 1980). This approach has since been expanded and applied in a variety of probabilistic settings (e.g., Clemen and Winkler, 1993; Kadane and Wolfson, 1996a, b).

4.4. Additional correlation-assessment concerns. Both the concordance probability assessment and the conditional fractile estimate may require that the expert be trained. For the concordance probability, training may include learning about probability assessment as well as clarification of the concordance event. For the conditional fractile estimate, the expert must understand fractiles. In addition, the expert must also understand the notion of regression toward the mean. It is evident from (6) that $|E[F(X) | y] - 0.5| \leq |G(y) - 0.5|$. Thus, in our example in which $G(180) = 0.70$, $E[F(X) | 180]$ must fall between 0.30 and 0.70.

On a technical issue, it is not just the assessment of the individual correlations that matters, but the assessment of \mathbf{R}_S overall; the analyst must ensure that \mathbf{R}_S is positive definite. If correlations are assessed individually without consideration of the overall joint distribution, it is quite possible (and in our experience surprisingly easy) to end up with a non-positive definite correlation matrix. If \mathbf{R}_S is poorly conditioned (analogous to multicollinearity in regression), this may lead to some very counterintuitive results; for example, conditional distributions and risk profiles may appear deceptively narrow, reflecting the mathematics of the specific model, even though such a result may be inconsistent with the expert's intuition.

One can imagine many different ways to adjust the assessed correlations if necessary to satisfy the constraint of positive definiteness, the ultimate goal being to obtain a correlation matrix that satisfies both the expert and the constraint. In this sense, the constraint can be viewed as ensuring coherence in the assessment of the dependency relationships and hence in the joint distribution. Research by the authors is under way to develop assessment procedures that will ensure that the assessed correlation matrix will be positive definite.

Each of the three methods for correlation assessment described above has something to recommend it. At this point in time, we can make no claim for the superiority of one method over

another regarding the cognitive appropriateness of the task, the ease of making the required judgment, or the accuracy of the correlations obtained. Research is currently under way by the authors and colleagues to study the effectiveness of these assessment approaches. For now, we recommend that analysts and experts use a combination of approaches in concert in order to find a correlation that satisfies the expert. For example, an initial assessment of Spearman's r can be used to generate a sample scatterplot based on simulated values. Likewise, if the analyst has adopted the multivariate normal copula, then r can be used to derive a corresponding P_C (via the expressions given above relating r , r , t , and P_C) and the expert can be asked if the implied P_C adequately reflects his or her reasoning. As usual, the assessment procedure would iterate until the expert is satisfied with all of the implications.

6. Discussion and Conclusion

We have argued that decision and risk analysis can benefit from the use of correlations and copulas in the construction of probabilistic models, and we have shown explicitly how to realize these benefits with practical correlation-assessment methods and through the use of the multivariate-normal copula. Throughout, our argument has been that this approach complements the set of tools available to the analyst, potentially streamlining the construction and analysis of probabilistic models. As with any modeling technique, however, there are a number of issues, including trade-offs that the analyst must make and further research needed to answer important questions. In this final section we discuss some of these issues.

With regard to the modeling process itself, there are many ways to construct and use copula-based joint densities that were not mentioned in Sections 2 and 3. For example, we focused on the multivariate normal copula, but others may also be useful, such as the block-

uniform family described by MacKenzie (1994). In some cases, decision analysts discretize density functions; we used the extended Pearson-Tukey discrete approximation method, but other methods may give better results. Modeling the entire joint distribution raises the question of how best to discretize a full joint distribution: How does one select appropriate representative points, and what probabilities should be applied? This question amounts to choosing representative scenarios in a complex multivariate space. Some efforts have been made in this direction (e.g., DeVuyst, Preckel, and Liu, 1996).

As mentioned in the introduction, copula-based modeling is most useful for situations involving continuous variables, but many important applications that use expert judgment, especially in artificial intelligence, involve the use of categorical variables, for which conditional reasoning may be more appropriate. Also, the marginal-and-conditional approach has the advantage that any plausible relationship can be modeled. The use of pairwise correlations, on the other hand, limits the nature of the relationships that can be modeled as discussed above in Section 2.2. In many cases, however, we believe that expert knowledge of relationships among variables may be adequately modeled in terms of pairwise dependence by using correlations.

The calculation of conditional distributions from the copula-based joint distribution is straightforward. One implication of this is that the search for conditional independence becomes less critical; it yields no real savings either in assessment or computational complexity. The ease of calculating conditional distributions will be especially useful for more complex models or analyses. For example, in a value-of-information analysis, it would be necessary to separate out one or more variables from the copula, conditioning decisions and the remaining copula variables on the information variable. For inference applications, the ease of calculating the conditional distributions should make it straightforward to propagate information through a network.

It is important to realize that good decision and risk analysis does incorporate thinking about relationships among variables. Conventionally, this is done through careful thinking about conditional relationships, often aided by causal reasoning on the part of the expert. This approach has proven itself as a useful way to cope with complex knowledge elicitation and modeling situations. Its primary drawback is that the assessment burden grows exponentially with the number of variables in the model. As we have demonstrated, use of correlations can potentially reduce the assessment burden. For many experts and many situations, thinking in terms of correlations among variables may be more natural than the conventional marginal-and-conditional approach.

Acknowledgments

We thank Greg Fischer, George MacKenzie, and Bob Winkler for many discussions on both general and technical issues relating to the assessment of correlations and the use of copulas in decision analysis. Jim Smith and Bob Winkler provided useful comments on various drafts of this paper. This work was supported by the National Science Foundation under Grant SBR 95-96176.

References

- Barlow, R. E., and F. Proschan, *Statistical theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart, and Winston, New York, 1975.
- Burns, W. J., and R. T. Clemen, "Covariance Structure Models and Influence Diagrams," *Management Science*, 39 (1993), 816-834.
- Clemen, R. T., and R. L. Winkler, "Aggregating Point Estimates: A Flexible Modeling Approach," *Management Science*, 39, (1993), 501-515.

- Clemen, R. T., *Making Hard Decisions: An Introduction to Decision Analysis*, 2nd Ed., Duxbury, Belmont, CA, 1996.
- Dall'Aglio, G., S. Kotz, and G. Salinetti, *Advances in Probability Distributions with Given Marginals*, Kluwer, Dordrecht, Netherlands, 1991.
- DeVuyst, E. A., P. V. Preckel, and S. Liu, "Discrete Approximations of Joint Probability Distributions," University of Illinois (1996), draft.
- Frank, M. J., "On the Simultaneous Associativity of $F(x, y)$ and $x + y - F(x, y)$," *Aequationes Mathematicae*, 19 (1979), 194-226.
- Gigerenzer, G., "How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases," *European Review of Social Psychology*, 2 (1991), 83-115.
- Gigerenzer, G., U. Hoffrage, and H. Kleinbolting, "Probabilistic Mental Models: A Brunswikian Theory of Confidence," *Psychological Review*, 98 (1991), 506-528.
- Gokhale, D. V., and S. J. Press, "Assessment of a Prior Distribution for the Correlation Coefficient in a Bivariate Normal Distribution," *Journal of the Royal Statistical Society (A)*, 145 Part 2 (1982), 237-249.
- Howard, R. A., and J. Matheson, *The Principles and Applications of Decision Analysis*, SDG, Inc., Menlo Park, CA, 1983.
- Iman, R. L., and W. J. Conover, "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables," *Communications in Statistics -- Simulation and Computation*, 11 (1982), 311-334.
- Jouini, M., and R. T. Clemen, "Copula Models for Aggregating Expert Opinions," *Operations Research*, 44 (1996), 444-457.

- Kadane, J. B., J. Dickey, R. L. Winkler, W. Smith, and S. Peters, "Interactive Elicitation of Opinion for a Normal Linear Model," *Journal of the American Statistical Association*, 75 (1980), 845-854.
- Kadane, J. B., and L. J. Wolfson, "Priors for the Design and Analysis of Clinical Trials," University of Waterloo, draft (1996a).
- Kadane, J. B., and L. J. Wolfson, "Experiences in Elicitation," University of Waterloo, draft (1996b).
- Keefer, D., and S. E. Bodily, "Three-Point Approximations for Continuous Random Variables," *Management Science*, 29 (1983), 595-609.
- Kelly, K. S., and R. Krzysztofowicz, "A Bayesian Model and Choice of Expert," University of Virginia (1996a), draft.
- Kelly, K. S., and R. Krzysztofowicz, "A Bayesian Model of Multiple Experts," University of Virginia (1996b), draft.
- Kruskal, W., "Ordinal Measures of Association," *Journal of the American Statistical Association*, 53 (1958), 814-861.
- Lehmann, E. L., "Some Concepts of Dependence," *Annals of Mathematical Statistics*, 37 (1966), 1137-1153.
- MacKenzie, G. R., *Approximately Maximum-Entropy Multivariate Distributions with Specified Marginals and Pairwise Correlations*, PhD Dissertation, University of Oregon (1994).
- Matheson, J. E., and R. A. Howard, "An Introduction to Decision Analysis," Stanford Research Institute, Palo Alto, CA, 1968. Reprinted in Howard and Matheson (Eds.), *The Principles and Applications of Decision Analysis*, SDG, Inc., Menlo Park, CA (1983), 17-55.
- Moder, J. J., and E. G. Rodgers, "Judgment Estimates of the Moments of PERT Type Distributions," *Management Science*, 15 (1968), B76-B83.

- Morgan, M. G., and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge, 1990.
- Nelsen, R. B., "Properties of a One-Parameter Family of Bivariate Distributions with Specified Marginals," *Communications in Statistics -- Theory and Methods*, 15 (1986), 3277-3285.
- Nelsen, R. B., "Copulas and Association," in G. Dall'Aglio, S. Kotz, and G. Salinetti (Eds.), *Advances in Probability Distributions with Given Marginals*, Kluwer, Dordrecht, Netherlands, 1991.
- Perry, C., and I. D. Grieg, "Estimating the Mean and Variance of Subjective Distributions in PERT and Decision Analysis," *Management Science*, 21 (1975), 1477-1480.
- Reilly, T., "Sensitivity Analysis for Dependent Variables," University of Oregon (1996), draft.
- Sklar, A., "Fonctions de Répartition à n Dimensions et Leurs Marges," *Publications de l'Institut Statistique de l'Université de Paris*, 8 (1959), 229-231.
- Winkler, R. L., W. Smith, and R. Kulkarni, "Adaptive Forecasting Models based on Predictive Distribution," *Management Science*, 24 (1978), 977-986.
- Yi, W., and V. M. Bier, "An Application of Copulas to Accident Precursor Analysis," University of Wisconsin, Madison (1996), draft.

Footnotes

- ¹ Although Spearman's ρ was originally developed as a measure of association in a sample, it has a population analog that can be expressed in terms of a copula. See, for example, Nelsen (1991).

² Strictly speaking, weaker definitions of dependence would work. For example, Lehmann defines *positive quadrant dependence* as follows: X_i and X_j are positive quadrant dependent if $P(X_i \leq x_i \text{ and } X_j \leq x_j) \geq P(X_i \leq x_i)P(X_j \leq x_j)$. Furthermore, Lehmann shows that positive regression dependence implies positive quadrant dependence, which in turn implies that $r \geq 0$. We prefer the stronger concept of regression dependence because its definition is intuitively grasped and is easily seen to be consistent with regression-like relationships.

³ The scaled beta density $f_b(x)$ is created by specifying a closed interval $[x_0, x_1]$ for the support of the random variable X and then calculating the density as

$$f_b(x | a, b, x_0, x_1) = \{[(x-x_0)/(x_1-x_0)]^{a-1} [(x_1-x)/(x_1-x_0)]^{b-1}\} / [B(a, b) (x_1-x_0)],$$

where $B(a, b)$ is the beta function.