Reliable Computing **10** (2) (2004), pp. 139-161.

Using Pearson Correlation to Improve Envelopes Around the Distributions of Functions

Daniel Berleant and Jianzhong Zhang
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011

Abstract

Given two random variables whose dependency relationship is unknown, if a new random variable is defined whose samples are some function of samples of the given random variables, the distribution of this function is not fully determined. However, envelopes can be computed that bound the space through which its cumulative distribution function must pass. If those envelopes could be made to bound a smaller space, the cumulative distribution, while still not fully determined, would at least be more constrained. We show how information about the correlation between values of given random variables can lead to better envelopes around the cumulative distribution of a function of their values.

1 Introduction and Background

A random variable whose samples are a function of samples of other random variables is often called a *derived random variable* and its distribution a *derived distribution*. Given two random variables with samples u and v, probability density functions $f_u(.)$ and $f_v(.)$, and cumulative distribution functions $F_u(.)$ and $F_v(.)$, a sample x of a derived distribution can be defined in various ways, such as:

- $x = u + v$ (Frank et al. 1987);
- $x = max(u, v)$, which models the time to complete two concurrent tasks; and
- $x_v = \frac{38u - 8v}{0.08u + 0.048v}$, where u and v are the fuel cost rates of two electric generators and x_v is the optimal power output of the generator with rate v (Wood and Wollenberg 1996).

We wish to describe the distribution $F_x(.)$ of x.

Derived distributions may be determined analytically or numerically. Analytical methods tend either to assume distributions are of particular forms or, in the case of moment propagation, to ignore other information about the distributions. Springer (1979 [18]) gives a reasonably comprehensive account up to its time of publication. We pursue the numerical strategy here. Our strategy represents each input probability density function (PDF) discretely using a histogram-like set of intervals with associated probabilities [2]. The discretized inputs form the marginals of a discretized joint distribution termed a *joint distribution tableau*. Each cell in a joint distribution tableau contains an interval and a probability, and is termed a marginal cell if it contains an

$\mathbf{v}_3 = (5, 9]$ $p(v \in \mathbf{v}_3) = 0.1$	$x = \frac{v}{u} \in \frac{\mathbf{v}_3}{\mathbf{u}_1} = (\frac{5}{2}, 9]$ $p_{13} = 0.02 =$ $p(u \in \mathbf{u}_1 \text{ and } v \in \mathbf{v}_3)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_3}{\mathbf{u}_2} = (\frac{5}{3}, \frac{9}{2})$ $p_{23} = 0.05 =$ $p(u \in \mathbf{u}_2 \text{ and } v \in \mathbf{v}_3)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_3}{\mathbf{u}_3} = (\frac{5}{4}, 3)$ $p_{33} = 0.03 =$ $p(u \in \mathbf{u}_3 \text{ and } v \in \mathbf{v}_3)$
$\mathbf{v}_2 = (4, 5]$ $p(v \in \mathbf{v}_2) = 0.8$	$x = \frac{v}{u} \in \frac{\mathbf{v}_2}{\mathbf{u}_1} = (2, 5]$ $p_{12} = 0.16 =$ $p(u \in \mathbf{u}_1 \text{ and } v \in \mathbf{v}_2)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_2}{\mathbf{u}_2} = (\frac{4}{3}, \frac{5}{2})$ $p_{22} = 0.4 =$ $p(u \in \mathbf{u}_2 \text{ and } v \in \mathbf{v}_2)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_2}{\mathbf{u}_3} = (1, \frac{5}{3})$ $p_{32} = 0.24 =$ $p(u \in \mathbf{u}_3 \text{ and } v \in \mathbf{v}_2)$
$\mathbf{v}_1 = [0, 4]$ $p(v \in \mathbf{v}_1) = 0.1$	$x = \frac{v}{u} \in \frac{\mathbf{v}_1}{\mathbf{u}_1} = [0, 4]$ $p_{11} = 0.02 =$ $p(u \in \mathbf{u}_1 \text{ and } v \in \mathbf{v}_1)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_1}{\mathbf{u}_2} = [0, 2)$ $p_{21} = 0.05 =$ $p(u \in \mathbf{u}_2 \text{ and } v \in \mathbf{v}_1)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_1}{\mathbf{u}_3} = [0, \frac{4}{3})$ $p_{31} = 0.03 =$ $p(u \in \mathbf{u}_3 \text{ and } v \in \mathbf{v}_1)$
$v \uparrow \quad x = \frac{v}{u} \nearrow$ $u \rightarrow$	$\mathbf{u}_1 = [1, 2]$ $p(u \in \mathbf{u}_1) = 0.2$	$\mathbf{u}_2 = (2, 3]$ $p(u \in \mathbf{u}_2) = 0.5$	$\mathbf{u}_3 = (3, 4]$ $p(u \in \mathbf{u}_3) = 0.3$

Table 1: a joint distribution tableau. Independent random variables with PDFs $f_u(.)$ and $f_v(.)$ are shown in discretized form, using intervals $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ and their associated probabilities to represent $f_u(.)$ in the bottom row, and intervals $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ and their associated probabilities to represent $f_v(.)$ in the left column. Values u and v are drawn from $f_u(.)$ and $f_v(.)$. The discretization is coarse for illustration. We have discretized $f_u(.)$ and $f_v(.)$ without overlaps, so some intervals have open endpoint(s). These are shown with a parenthesis instead of a square bracket.

interval $\mathbf{u}_i$ or $\mathbf{v}_j$ in the discretization of marginal $f_u(.)$ or $f_v(.)$, and an interior cell if contains an interval and a probability p_{ij} (Table 1). For each i, j, $p_{ij} = p(u \in \mathbf{u}_i \text{ and } v \in \mathbf{v}_j)$. If the inputs are statistically independent in the usual sense then $p(u \in \mathbf{u}_i \text{ and } v \in \mathbf{v}_j) = p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_j)$.

If each probability p_{ij} is assumed to be distributed uniformly over its corresponding interval $\mathbf{v}_j / \mathbf{u}_i$, a not unreasonable approximation if the discretization is sufficiently fine, then the cumulative distribution function (CDF) of x, call it $F_x(x_0)$, could be plotted by taking values x_0 and performing the following steps for each (Moore 1984).

1. Integrate each interior cell from $-\infty$ to x_0.
2. Sum the integrals computed for the interior cells.

The PDF $f_x(.)$ instead of the CDF $F_x(.)$ can also be obtained (Ingram et al. 1968; Colombo and Jaarsma 1980). If no assumption is made about the distribution of the p_{ij}'s over their respective domains, then $F_x(.)$ cannot be determined precisely, but can be bounded with envelopes (Figure 1) which bound the effects of discretization [2].

A problem with such methods is the need to know the dependency relationship between the input distributions. Independence is a common assumption in practice though not always justified. Independence as well as other dependency relationships (as in Table 2) can be represented in a joint distribution tableau by appropriate choice of interior cell probabilities. However, sometimes no specification of dependency is justified by what is known about the problem.

There are a number of approaches to the problem of numerically computing derived distributions without specifying a dependency relationship between the operands (Figure 2 shows an example). One is Monte Carlo simulation (MC), as in Red-Horse and Benjamin [16]. However the randomness inherent in MC can lead to complications in the results and their interpretation (Ferson 1996). Another approach is based on copulas (Frank et al. 1987; Nelsen 1999), and a tool implementing the Probabilistic Arithmetic (Williamson and Downs 1990) extension of Frank et al. is available commercially (Ferson 2002). A third approach, clouds, was recently

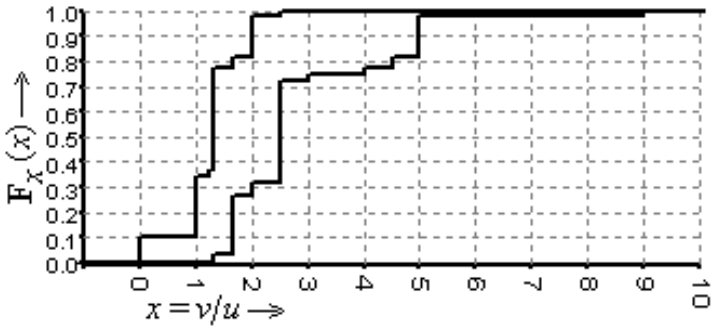

Figure 1: envelopes $\mathbf{F}_x(x)$ around CDF $F_x(x)$, where $x = v/u$ and $f_u(.)$ and $f_v(.)$ are discretized as shown in Table 1.

$\mathbf{v}_3 = (5, 9]$ $p(v \in \mathbf{v}_3) = 0.1$	$x = \frac{v}{u} \in \frac{\mathbf{v}_3}{\mathbf{u}_1} = (\frac{5}{2}, 9]$ $p_{13} = 0.1 =$ $p(u \in \mathbf{u}_1 \text{ and } v \in \mathbf{v}_3)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_3}{\mathbf{u}_2} = (\frac{5}{3}, \frac{9}{2})$ $p_{23} = 0 =$ $p(u \in \mathbf{u}_2 \text{ and } v \in \mathbf{v}_3)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_3}{\mathbf{u}_3} = (\frac{5}{4}, 3)$ $p_{33} = 0 =$ $p(u \in \mathbf{u}_3 \text{ and } v \in \mathbf{v}_3)$
$\mathbf{v}_2 = (4, 5]$ $p(v \in \mathbf{v}_2) = 0.8$	$x = \frac{v}{u} \in \frac{\mathbf{v}_2}{\mathbf{u}_1} = (2, 5]$ $p_{12} = 0 =$ $p(u \in \mathbf{u}_1 \text{ and } v \in \mathbf{v}_2)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_2}{\mathbf{u}_2} = (\frac{4}{3}, \frac{5}{2})$ $p_{22} = 0.5 =$ $p(u \in \mathbf{u}_2 \text{ and } v \in \mathbf{v}_2)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_2}{\mathbf{u}_3} = (1, \frac{5}{3})$ $p_{32} = 0.3 =$ $p(u \in \mathbf{u}_3 \text{ and } v \in \mathbf{v}_2)$
$\mathbf{v}_1 = [0, 4]$ $p(v \in \mathbf{v}_1) = 0.1$	$x = \frac{v}{u} \in \frac{\mathbf{v}_1}{\mathbf{u}_1} = [0, 4]$ $p_{11} = 0.1 =$ $p(u \in \mathbf{u}_1 \text{ and } v \in \mathbf{v}_1)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_1}{\mathbf{u}_2} = [0, 2)$ $p_{21} = 0 =$ $p(u \in \mathbf{u}_2 \text{ and } v \in \mathbf{v}_1)$	$x = \frac{v}{u} \in \frac{\mathbf{v}_1}{\mathbf{u}_3} = [0, \frac{4}{3})$ $p_{31} = 0 =$ $p(u \in \mathbf{u}_3 \text{ and } v \in \mathbf{v}_1)$
$v \uparrow \quad x = \frac{v}{u} \nearrow$ $u \rightarrow$	$\mathbf{u}_1 = [1, 2]$ $p(u \in \mathbf{u}_1) = 0.2$	$\mathbf{u}_2 = (2, 3]$ $p(u \in \mathbf{u}_2) = 0.5$	$\mathbf{u}_3 = (3, 4]$ $p(u \in \mathbf{u}_3) = 0.3$

Table 2: a joint distribution tableau like that of Table 1 except with different values for the p_{ij}'s, indicating that the joint distribution is different. Hence the dependency relationship between values u and v of the marginals is also different.

proposed by Neumaier [15]. A fourth approach is discrete convolution of the actual distributions. Various techniques based on this approach have existed since at least as early as 1968 [12] for the case of independence. More recently, the technique described here, called Distribution Envelope Determination (DEnv), extended the discrete convolution technique to the case of unknown dependency (Berleant and Goodman-Strauss 1998 [3]). DEnv is reviewed in the next section. It is implemented in a downloadable tool called Statool [19].

Finally, the intermediate situation of *partial information* about the dependency may occur. There is a need for ways to use partial information about dependency between inputs when determining envelopes around the CDFs of derived distributions [10]. A common and important way to express partial information about dependency is correlation. Correlation constitutes partial information because it does not fully characterize a dependency relationship (different joint distributions can have exactly the same correlation). We have extended DEnv to incorporate information about correlation. We use Pearson correlation, the most common kind and the kind normally implied by uses of the otherwise ambiguous term "correlation." (In copula-based approaches, handling Pearson correlation is problematic [20] because converting joint distributions into copulas involves stretching the marginals into a normalized form, and Pearson correlation depends on the un-normalized forms.) The purpose of this paper is to report on an extension of DEnv that uses Pearson correlation as a problem input.

We review the DEnv algorithm next (a more detailed account appears in [5]). Then we explain how to extend DEnv to use correlation to provide constraints that can often decrease the separation of the envelopes.

2 Distribution Envelope Determination (DEnv): a Review

The goal. DEnv obtains boundaries around the space through which a derived CDF may travel (Figure 2). More specifically, let $F_x(.)$ be the cumulative distribution for x, where x is a function of u and v. The density function $f_u(.)$ of u is discretized with a set of intervals $\mathbf{u}_i$, each associated with a probability such that the sum of these probabilities is 1. Density function $f_v(.)$ of v is similarly discretized with a set of intervals $\mathbf{v}_j$. Because the discretizations lose information that is present in the undiscretized $f_u(.)$ and $f_v(.)$, there will typically not be a single CDF that is implied for $x = g(u, v)$ even when the dependency relationship is fully specified [2]. Our objective then is to obtain left and right envelopes around the family of CDFs that are possible for x. These envelopes may be expressed symbolically as the interval-valued function $\mathbf{F}_x(.)$. The left (top) envelope then is $\overline{\mathbf{F}_x}(.)$ and the right (bottom) envelope is $\underline{\mathbf{F}_x}(.)$

The givens. Envelope computation takes as input the correlation between the marginals, when that is available, and a joint distribution tableau. A joint distribution tableau discretely represents a family of joint distributions containing all joint distributions that are consistent with that discretization. For example, recall the joint distribution tableau of Table 1. This tableau states that $p(v \in [0, 4]) = 0.1$, $p(u \in [1, 2]) = 0.2$, and $p(v \in [0, 4]$ and $u \in [1, 2]) = 0.02$. Each cell in the tableau contains an interval-valued bin in which u or v (for a marginal cell), or $x = v/u$ (for an interior cell) might fall, and a probability that it falls in that bin. The probabilities of interior cells are specified if the dependency relationship of the marginals is known, and not specified if the dependency is not known. There are many variations in how values u and v of the marginals can be distributed, and in how they can be jointly distributed, that are consistent with these bin specifications. Put another way, Table 1 gives a correct discretization of any pair of marginal distributions and their joint distribution for which the statements in all of the cells are correct. Table 1 also contains a discretization of the distribution of x. This is the set of interior cells, each of which specifies an interval-valued bin for $x = v/u$ and a probability p_{ij}.

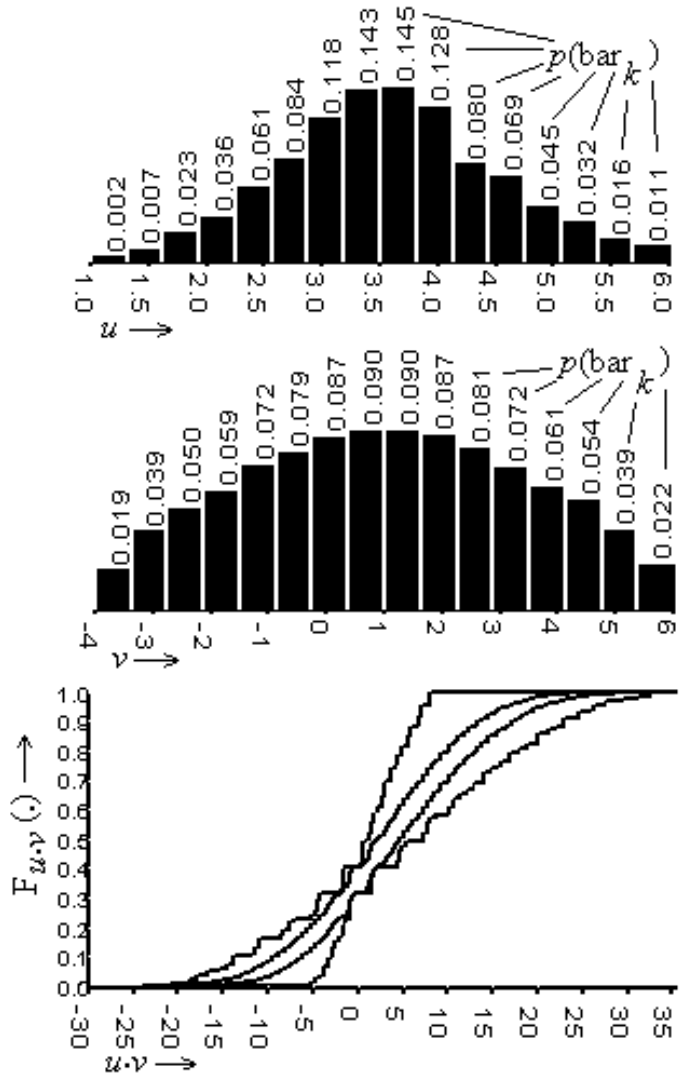

Figure 2: (*top and middle*) histogram-like discretizations of input PDFs $f_u(u)$ and $f_v(v)$. Each bar is labeled at the top with its probability. (*Bottom*) two pairs of envelopes around $F_{u \cdot v}(.)$, the CDF of derived value $x = u \cdot v$. The two exterior envelopes bound the CDF when the dependency relationship between u and v is unknown. The two interior envelopes bound the CDF when u and v are independent. In the independent case, the envelopes are non-identical because they bound the effects of information loss due to discretization. The rougher appearance of both pairs of envelopes near $u \cdot v = 0$ is because $0 \cdot \text{anything} = 0$.

In the following subsections we first assume independence in the traditional sense (Section 2.1), then extend that to arbitrary dependency relationships (Section 2.2), then further the algorithm to the case of an unknown dependency relationship (Section 2.3). With that as background the case of a dependency relationship constrained by correlation is finally addressed (Section 3). That case constitutes the new contribution of this report.

2.1 Solution for Independent Marginals

Equations (1)-(2) summarize the solution for the general case of $x = g(u, v)$, with interval extension $\mathbf{x}_{ij} = \mathbf{g}(\mathbf{u}_i, \mathbf{v}_j)$ where $\mathbf{u}_i$ and $\mathbf{v}_j$ are intervals in discretizations of the distributions $f_u(.)$ and $f_v(.)$ from which values u and v are drawn.

$$\overline{\mathbf{F}_x}(x_0) = \sum_{i,j:\underline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0} p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_j) \tag{1}$$

$$\underline{\mathbf{F}_x}(x_0) = \sum_{i,j:\overline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0} p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_j) \tag{2}$$

The summations are over all pairs i, j such that $\underline{\mathbf{g}(\mathbf{u}_i, \mathbf{v}_j)} \le x_0$ in Equation (1), or $\overline{\mathbf{g}(\mathbf{u}_i, \mathbf{v}_j)} \le x_0$ in Equation (2).

We first explain why Equation (1) computes the left bounding envelope $\overline{\mathbf{F}_x}(x)$, using an example. Then the differences for the right envelope are noted. Bolding will indicate an interval and overlining the upper bound of an interval.

2.1.1 Computing the Left (Upper) Envelope from a Joint Distribution Tableau.

The example is based on Table 1 and is stated in several steps.

1. For $x < 0$, $\overline{\mathbf{F}_x}(x) = 0$ because no interior cell contains an interval containing any values below zero, so $p(x < 0)$ must be zero.
2. For $0 \le x \le 1$, $\overline{\mathbf{F}_x}(x) = p_{11} + p_{21} + p_{31} = 0.1$ for the following reasons.
   - Only the interior cells containing p_{11}, p_{21}, and p_{31} have intervals with low bounds ≤ 1. Therefore only those cells can contain x when $x \le 1$, thereby contributing their probability to the cumulative probability $\overline{\mathbf{F}_x}(x)$.
   - Call the distribution of a particular interior cell's probability over its interval its *mini-distribution.* The probability associated with an interior cell must be distributed somehow within its interval, but mini-distributions are not otherwise defined. Therefore to obtain the height of the left bounding envelope at a given value of x we must assume that each mini-distribution has a form that leads to the greatest possible height at that value. The simplest such assumption is that the mini-distribution of each interior cell interval is an impulse at its low bound, because then each interior cell whose interval low bound is at or below a value $x = x_0$ will contribute all of its probability to $\overline{\mathbf{F}_x}(x_0)$.
3. For $1 < x \le \frac{5}{4}$, $p_{32} = 0.24$ can also contribute to $\overline{\mathbf{F}_x}(x)$, so $\overline{\mathbf{F}_x}(x) = p_{11} + p_{21} + p_{31} + p_{32} = 0.34$.
4. For $\frac{5}{4} < x \le \frac{4}{3}$, $p_{33} = 0.03$ can also contribute to $\overline{\mathbf{F}_x}(x)$, for a total cumulative probability of 0.37.
5. This line of reasoning continues until all interior cells contribute their probabilities to $\overline{\mathbf{F}_x}(x)$, resulting in the staircase-shaped left envelope shown in Figure 1.

2.1.2 Computing the Right (Lower) Envelope

The right bounding envelope, $\underline{\mathbf{F}_x}(x)$, is derived similarly, except that points on it are obtained by assuming that the probability in each interior cell is an impulse at its interval high bound instead of its interval low bound.

2.2 Solution for an Arbitrary Dependency Between the Marginals

In Table 1, each $p_{ij} = p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_j)$ is the product of the probabilities of its corresponding marginal cells. This is consistent with the traditional definition of statistical independence. Other assignments of probabilities to the p_{ij}'s imply other dependency relationships. If the dependency relationship is known then the interior cells can be filled in so that their probabilities are consistent with that dependency relationship and its joint distribution. In such cases the value of each p_{ij} is not necessarily $p(u \in \mathbf{u}_i) \cdot p(v \in \mathbf{v}_v)$, instead arising out of the dependency relationship, which defines the value of $p(u \in \mathbf{u}_i \text{ and } v \in \mathbf{v}_j)$. This implies a generalization of Equations (1)-(2), shown as Equations (3)-(4).

$$\overline{\mathbf{F}_x}(x_0) = \sum_{i,j:\underline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0} p(u \in \mathbf{u}_i \text{ and } v \in \mathbf{v}_j) \tag{3}$$

$$\underline{\mathbf{F}_x}(x_0) = \sum_{i,j:\overline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0} p(u \in \mathbf{u}_i \text{ and } v \in \mathbf{v}_j) \tag{4}$$

2.3 Solution for the Case of Unknown Dependency Between the Marginals

As explained earlier (Section 2), the interior cells of a joint distribution tableau represent a family of CDFs. When the dependency relationship between the marginals is unknown, then Equations (1)-(4) cannot be evaluated because the p_{ij}s are not determined. Intuitively, because the p_{ij}'s are now variable, they may take on values consistent with a greater variety of joint distributions, and hence a greater variety of CDFs for derived random variable x. This will tend to make the envelopes bounding this larger family of CDFs wider apart. An augmentation to the algorithm is required to deal with this situation. The augmented algorithm is described next in two steps, one short and one longer, and then summarized in Equations (5)-(9).

1. *Determine which interior cells contribute.* The same cells contribute their probabilities to the CDF at a value of x as would contribute in the case of known dependency, and for the same reasons. These are the cells specified by Equations (1)-(4).

2. *Maximize (for the left envelope), or minimize (for the right envelope) the sum of the probabilities of the contributing cells.* Because the p_{ij}'s are not fully determined when the dependency relationship is unknown, DEnv finds maximums and minimums given the result of step 1 by manipulating the p_{ij}'s in the joint distribution tableau. Call the interior cells identified in step 1 the *contributing cells,* and the cells containing the remaining p_{ij}'s the *non-contributing cells.*

   To maximize the sum of the probabilities of the contributing cells, we transfer as much probability as possible from non-contributing interior cells to contributing interior cells. To illustrate, recall Table 1. If the assumption that the marginals are independent is relaxed, the p_{ij}'s are underdetermined. However they are constrained by the fact that the probabilities of the interior cells in any given row must sum to the probability of its corresponding marginal cell, and similarly for any given column (Table 3).

   For example, compare the assignment of probability to p_{32} in Table 2 with its assignment in Table 1, 0.3 vs. 0.24. In Table 2, $\overline{\mathbf{F}_x}(1.1) = p_{11} + p_{21} + p_{31} + p_{32} = 0.4$, which is greater

$p(v \in \mathbf{v}_3) = 0.1$	p_{13}	p_{23}	p_{33}
$p(v \in \mathbf{v}_2) = 0.8$	p_{12}	p_{22}	p_{32}
$p(v \in \mathbf{v}_1) = 0.1$	p_{11}	p_{21}	p_{31}
$v \uparrow \quad u \rightarrow$	$p(u \in \mathbf{u}_1) = 0.2$	$p(u \in \mathbf{u}_2) = 0.5$	$p(u \in \mathbf{u}_3) = 0.3$

Row constraints	**Column constraints**
$p_{11} + p_{21} + p_{31} = 0.1$	$p_{11} + p_{12} + p_{13} = 0.2$
$p_{12} + p_{22} + p_{32} = 0.8$	$p_{21} + p_{22} + p_{23} = 0.5$
$p_{13} + p_{23} + p_{33} = 0.1$	$p_{31} + p_{32} + p_{33} = 0.3$

Table 3: (*top*) a joint distribution tableau like that of Tables 1 and 2, but showing only the p_{ij}'s and without values assigned to them. (*Bottom*) the constraints that the tableau defines on the values of the p_{ij}'s. Each constraint states that the sum of the probabilities of the p_{ij}'s in a row or column equals the probability in the marginal cell for that row or column. This follows from standard properties of joint distributions and their marginals.

than the 0.34 implied by $p_{11}+p_{21}+p_{31}+p_{32}$ in Table 1. $\overline{\mathbf{F}_x}(1.1)$ can be no higher than the Table 2 value of 0.4 no matter what the joint distribution is, because the third row must comply with the constraint $p_{11} + p_{21} + p_{31} = p(v \in \mathbf{v}_1) = 0.1$, and the only contributing interior cell outside of the third row is the one containing p_{32}, which can be no higher than 0.3 because its column must comply with the constraint $p_{31} + p_{32} + p_{33} = p(u \in \mathbf{u}_3) = 0.3$. The result is a point on the left envelope at $x = 1.1$ that is higher than the envelope derived for the independent case, a new height that applies not only to $x = v/u = 1.1$ but also to all values of $x = v/u$ for which the contributing cells are p_{11}, p_{21}, p_{31}, and p_{32}. For other values of x the set of contributing cells is different, so the p_{ij}'s of Table 2 might not lead to the highest possible value of $\overline{\mathbf{F}_x}(x)$. In that case some other set of assignments of probabilities to the p_{ij}'s consistent with Table 3 will result in the highest possible value instead. Thus for each value of $x = v/u$ it is necessary to find the contributing cells, and assignments to the p_{ij}'s in them that lead to the highest possible value of $\overline{\mathbf{F}_x}(x)$. The result is ultimately a left envelope that is farther to the left than the left envelope shown in Figure 1. Similar reasoning based on minimization instead of maximization gives a new right envelope that is farther to the right than the one shown in Figure 1.

Maximizing the collective probability of a set of contributing cells by the ad hoc reasoning process used for $x = 1.1$ for various values of x would rapidly become tedious to do manually. Fortunately a general and automatable method is available in the form of linear programming (LP). LP optimizes (maximizes or minimizes) a linear function, called the objective function, with respect to a set of linear constraints. The linear function to optimize in this case is the sum of the probabilities of the contributing cells. LP will maximize this consistently with the linear constraints imposed by the marginals, one constraint for each $\mathbf{u}_i$ and one for each $\mathbf{v}_j$ in the joint distribution tableau (Table 3). LP is invoked and its output, the maximum (minimum) possible total probability that can be allocated among the contributing cells, is the y coordinate associated with x, thus completing the coordinates for a point on the left (right) envelope.

The extensions of Equations (1)-(2) and (3)-(4) to objective functions to optimize for the unknown dependency situation are:

$$\overline{\mathbf{F}_x}(x_0) = \max \sum_{i,j:\underline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0} p_{ij} \tag{5}$$

for the left envelope, and

$$\underline{\mathbf{F}_x}(x_0) = \min \sum_{i,j:\overline{\mathbf{g}(\mathbf{u}_i,\mathbf{v}_j)} \le x_0} p_{ij} \tag{6}$$

for the right envelope. The applicable constraints are:

$$\sum_j p_{ij} = p(\mathbf{u}_i), \qquad \text{for all } i \tag{7}$$

$$\sum_i p_{ij} = p(\mathbf{v}_j), \qquad \text{for all } j \tag{8}$$

$$p_{ij} \ge 0, \qquad \text{for all } i, j. \tag{9}$$

3 Using Correlation to Move the Envelopes Closer Together

Specifying a dependency relationship between the input random variables implies envelopes that are closer together than when the dependency is unknown (Figure 2). A value or range for correlation is a *partial* specification of the dependency, and so implies envelopes that are:

1) at least as close together as when the dependency is unknown, but
2) at least as far apart as when the dependency is fully specified.

DEnv infers the effects of constraints on envelopes via calls to a linear programming routine. Thus to use information about correlation, this information must be expressed as linear constraints. These constraints can then supplement the row and column constraints used by the LP calls. This is explained next, while Section 4 provides examples.

We begin with a standard formula for the Pearson correlation coefficient ρ. We use Pearson correlation in this paper as it is the most common kind of correlation and is usually implied by otherwise unqualified uses of the term "correlation."

$$\rho = \frac{E(uv) - E(u)E(v)}{\sqrt{[E(u^2) - E(u)^2][E(v^2) - E(v)^2]}} = \frac{\mu_{u \cdot v} - \mu_u \cdot \mu_v}{\sqrt{\sigma_u^2 \cdot \sigma_v^2}}. \tag{10}$$

Here ρ is the Pearson correlation coefficient of the distributions of u and v, u and v are values to be drawn from the marginal distributions, $E(u)$ is the expectation function and is equivalent to the mean μ_u, $E(u^2) - E(u)^2 = \sigma_u^2$ is the variance of u, and similarly for v. Since ρ and the marginals are problem inputs, all terms can be computed from the inputs except $E(uv)$, the only term that depends on the joint distribution. Solving for $E(uv)$ gives

$$E(uv) = E(u)E(v) + \rho\sqrt{[E(u^2) - E(u)^2][E(v^2) - E(v)^2]}. \tag{11}$$

Because DEnv uses the PDFs of u and v after they have been discretized into sets of intervals and their associated probabilities, and because the distribution of each associated probability over its interval is unspecified, terms in Equation (11) can be determined only to within intervals. For example, given the discretized distribution of v in Tables 1 and 2,

$$E(v) \in 0.1 \cdot [0, 4] + 0.8 \cdot (4, 5] + 0.1 \cdot (5, 9] = (3.7, 5.3]. \tag{12}$$

If we follow the convention of bolding interval-valued symbols, then $\mathbf{E}(v) = (3.7, 5.3]$. This leads to an intervalized form of Equation (11) suitable for use with discrete representations of PDFs for u and v.

$$\mathbf{E_g} = \mathbf{E}(uv) = \mathbf{E}(u)\mathbf{E}(v) + \boldsymbol{\rho}\sqrt{[\mathbf{E}(u^2) - \mathbf{E}(u)^2][\mathbf{E}(v^2) - \mathbf{E}(v)^2]} = \boldsymbol{\mu}_u\boldsymbol{\mu}_v + \boldsymbol{\rho}\sqrt{\boldsymbol{\sigma}_u^2\boldsymbol{\sigma}_v^2} \qquad (13)$$

Thus $\mathbf{E}(uv)$ is calculated from $\boldsymbol{\rho}$ and discretizations of the PDFs of u and v. Since $\boldsymbol{\rho}$ and the marginals are **g**ivens, we will call this expectation $\mathbf{E_g}$.

Another way to calculate $E(uv)$ is directly from a joint distribution tableau. This gives an interval for $E(uv)$, namely $\sum_{i,j} \mathbf{u}_i\mathbf{v}_j p_{ij}$. See Table 4. Because it is computed as a property of the joint distribution, as expressed discretely by a given joint distribution **t**ableau, call it $\mathbf{E_t}(.)$. Its argument is a joint distribution tableau with a fully specified set of value assignments to the p_{ij}'s. The assignment of probability values to the interior cells of the joint distribution tableau, in conjunction with the $\mathbf{u}_i\mathbf{v}_j$ intervals, implies an interval for $\mathbf{E_t}(.)$ that must be consistent with $\mathbf{E_g}$ (which represents the discretized distributions of u and v and the given correlation). If $\mathbf{E_t}(.)$ and $\mathbf{E_g}$ are not consistent with each other, that assignment of values to the p_{ij}'s is not consistent with the given correlation and therefore is not allowed. As the following steps show, consistency means that $\mathbf{E_t}(.)$ and $\mathbf{E_g}$ overlap.

1. $\mathbf{E_g}$ is the interval of admissible values for $E(uv)$ based on $\boldsymbol{\rho}$ and other problem inputs as specified in Equation (13). The terms in (13) are all calculated from the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s (see e.g. Equation (12)). Because the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s appear repeatedly in (13), naïve interval evaluation will often result in $\mathbf{E_g}$ containing excess width, thereby weakening the power of $\mathbf{E_g}$ as a constraint on admissible values of $E(uv)$. To avoid that, an optimization technique can be used to compute good bounds for $\mathbf{E_g}$. Alternatively, values or ranges for the means (μ_u and μ_v) and variances (σ_u^2 and σ_v^2) of the marginals can be provided as problem inputs. This has the added benefit of allowing incorporation of mean and variance information that may be available and more specific than the bounds for mean and variance derivable directly from the discretized marginals.

2. $\mathbf{E_t}(.)$, in contrast to $\mathbf{E_g}$, is affected by the p_{ij}'s, which are determined by the joint distribution. An expression for $\mathbf{E_t}(.)$ may be derived as follows.

$$\begin{aligned}
\overline{\mathbf{E_t}(.)} &= \overline{\sum_{i,j} \mathbf{u}_i\mathbf{v}_j p_{ij}} && \text{(by the definition of expectation)} \\
&= \sum_{i,j} \overline{\mathbf{u}_i\mathbf{v}_j p_{ij}} && \text{(because the maximum of the sum is the sum of the maximums)} \\
&= \sum_{i,j} \overline{\mathbf{u}_i\mathbf{v}_j} p_{ij} && \text{(because the } p_{ij} \text{ are numbers)} \\
\underline{\mathbf{E_t}(.)} &= \sum_{i,j} \underline{\mathbf{u}_i\mathbf{v}_j} p_{ij} && \text{(similarly).}
\end{aligned} \qquad (14)$$

   To compute bounds on $\mathbf{E_t}(.)$ using Equations (14), the numerical value of each $\underline{\mathbf{u}_i\mathbf{v}_j}$ and $\overline{\mathbf{u}_i\mathbf{v}_j}$ term is needed. The standard definition of interval multiplication accounts for all possible combinations of signs on the bounds of $\mathbf{u}_i$ and $\mathbf{v}_j$ by multiplying each bound of $\mathbf{u}_i$ by each bound of $\mathbf{v}_j$ (four combinations), and using the *min* and *max* of the four as $\underline{\mathbf{u}_i\mathbf{v}_j}$ and $\overline{\mathbf{u}_i\mathbf{v}_j}$ respectively (e.g. Alefeld and Herzberger 1983).

3. The p_{ij}'s are variables because they are under-determined by the row and column constraints (Table 3). Assigning a specific set of values to the p_{ij}'s implies an associated

$\vdots$	$\ddots$	$\vdots$	$\iddots$
$\mathbf{v}_j$	$\cdots$	$x = g(u,v) \in \mathbf{g}(\mathbf{u}_i, \mathbf{v}_j) = \mathbf{x}_{ij}$ $p_{ij} = p(u \in \mathbf{u}_i \text{ and } v \in \mathbf{v}_j)$ $\mathbf{u}_i\mathbf{v}_j = [min(\underline{\mathbf{u}_i}\,\underline{\mathbf{v}_j},\ \underline{\mathbf{u}_i}\,\overline{\mathbf{v}_j},\ \overline{\mathbf{u}_i}\,\underline{\mathbf{v}_j},\ \overline{\mathbf{u}_i}\,\overline{\mathbf{v}_j}),$ $max(\underline{\mathbf{u}_i}\,\underline{\mathbf{v}_j},\ \underline{\mathbf{u}_i}\,\overline{\mathbf{v}_j},\ \overline{\mathbf{u}_i}\,\underline{\mathbf{v}_j},\ \overline{\mathbf{u}_i}\,\overline{\mathbf{v}_j})]$	$\cdots$
$\vdots$	$\iddots$	$\vdots$	$\ddots$
$v \uparrow \quad x = g(u,v) \nearrow$ $u \rightarrow$	$\cdots$	$\mathbf{u}_i$	$\cdots$

Table 4: abstract template for joint distribution tableaus. The bottom row includes a marginal cell describing the case where a value u drawn from marginal $f_u(.)$ falls within interval $\mathbf{u}_i$ of the discretization of $f_u(.)$. The left column includes a similar cell for v, $f_v(.)$, and $\mathbf{v}_j$. The function for combining values u and v is $g(u,v) = x$, its interval extension is $\mathbf{g}(\mathbf{u}_i, \mathbf{v}_j) = \mathbf{x}_{ij}$, and the distribution of value $x = g(u,v)$ is represented discretely by the interior cells of the tableau, one of which is shown in detail. Product $\mathbf{u}_i\mathbf{v}_j$ is used in calculating $\mathbf{E_t}(.)$, which is the range of possible values of $E(uv)$ for the different joint distributions consistent with the intervals and p_{ij}'s in the interior cells of the tableau.

interval $\mathbf{E_t}(.)$, which can be calculated per Equations (14). Some sets of value assignments to the p_{ij}'s imply intervals for $\mathbf{E_t}(.)$ that do not overlap $\mathbf{E_g}$. Those assignments are inconsistent with the correlation provided as a problem input (as explained in detail in the next step), and so can be excluded as implausible. Excluding a set of assignments to the p_{ij}'s can move the left envelope toward the right of where it would be if there was no information about correlation, and/or move the right envelope toward the left, narrowing their separation. This is because the excluded set of assignments might have a higher maximum cumulation $\overline{\mathbf{F}_x}(x)$ or lower minimum cumulation $\underline{\mathbf{F}_x}(x)$ for a given value of x than any that are not excluded.

4. The previous step stated that $\mathbf{E_t}(.)$ and $\mathbf{E_g}$ are inconsistent when they have no overlap. This step explains why. Specifying the values of the p_{ij}'s does not define the distribution of any p_{ij} over $\mathbf{u}_i\mathbf{v}_j$. Hence a joint distribution tableau with specified values for its p_{ij}'s represents a *family* of joint distributions. All joint distributions that conform to the discretization expressed by the joint distribution tableau are in that family.

   A joint distribution for values u and v has a numerical value for $E(uv)$. $\mathbf{E_t}(.) = \sum_{i,j} \mathbf{u}_i\mathbf{v}_j p_{ij}$ thus gives the range of numerical values for $E(uv)$ exhibited by the various joint distributions in the family associated with a particular set of value assignments to the p_{ij}'s. If $\mathbf{E_t}(.)$ does not intersect $\mathbf{E_g}$, then there is no joint distribution in that family for which $E(uv) \in \mathbf{E_g}$, so that set of value assignments to the p_{ij}'s is excludable as inconsistent with the value ρ or range $\boldsymbol{\rho}$ provided as a problem input. This requirement that $\mathbf{E_t}(.)$ and $\mathbf{E_g}$ overlap is stated in inequality form as the following two constraints:

$$\begin{aligned} \underline{\mathbf{E_g}} &\leq \overline{\mathbf{E_t}(.)} \quad \text{and} \\ \overline{\mathbf{E_g}} &\geq \underline{\mathbf{E_t}(.)}. \end{aligned} \tag{15}$$

5. To use constraints (15) in a linear programming problem, symbols $\underline{\mathbf{E_g}}$ and $\overline{\mathbf{E_g}}$ are replaced with their numerical values as calculated in step 1. $\underline{\mathbf{E_t}(.)}$ and $\overline{\mathbf{E_t}(.)}$ are replaced

with $\sum_{i,j} \underline{\mathbf{u}_i \mathbf{v}_j} p_{ij}$ and $\sum_{i,j} \overline{\mathbf{u}_i \mathbf{v}_j} p_{ij}$ respectively, as described in step 2. This results in Equations (16).

$$\begin{aligned} \underline{\boldsymbol{\mu}_u \boldsymbol{\mu}_v + \boldsymbol{\rho}\sqrt{\boldsymbol{\sigma}_u^2 \boldsymbol{\sigma}_v^2}} &\leq \sum_{i,j} \overline{\mathbf{u}_i \mathbf{v}_j} p_{ij} \quad \text{and} \\ \overline{\boldsymbol{\mu}_u \boldsymbol{\mu}_v + \boldsymbol{\rho}\sqrt{\boldsymbol{\sigma}_u^2 \boldsymbol{\sigma}_v^2}} &\geq \sum_{i,j} \underline{\mathbf{u}_i \mathbf{v}_j} p_{ij}. \end{aligned} \tag{16}$$

Since the only variables in Equations (16) are the p_{ij}'s, (16) constitutes linear constraints as required by LP. These can supplement the row and column constraints (Table 3), and will tend to result in envelopes that are closer together than those resulting from the row and column constraints alone.

3.1 Strengthening the Effect of Correlation

The width of interval $\mathbf{E_t}(.) = \sum_{i,j} \mathbf{u}_i \mathbf{v}_j p_{ij}$ is derived from the widths of the $\mathbf{u}_i \mathbf{v}_j$ terms. However if the distribution of each probability p_{ij} over the corresponding interval $\mathbf{u}_i \mathbf{v}_j$ was fully defined then the overall distribution of uv would be fully defined. Then a numerically-valued function, call it $E_t(.)$, could be calculated instead of the interval-valued function $\mathbf{E_t}(.)$. To define the distribution of each p_{ij} one might consider assuming that, as examples, the distribution of each p_{ij} over the interval $\mathbf{u}_i \mathbf{v}_j$ is uniform, or is an impulse at the midpoint of $\mathbf{u}_i \mathbf{v}_j$, or has some other fully defined form.

Since $E_t(.)$ is a number it will be narrower than the interval $\mathbf{E_t}(.)$, unless $\mathbf{E_t}(.)$ is a thin interval containing only one number. (This will occur in the important special case where u and v are discretized as series of impulses.) Suppose $E_t(.)$ is in fact narrower. Then it is less likely to intersect with $\mathbf{E_g}$ and so more likely to be excluded as inconsistent with $\mathbf{E_g}$. Thus constraints (15) would be strengthened, leading to envelopes that are closer together.

For example, assume the distribution of each p_{ij} is uniform over $\mathbf{u}_i \mathbf{v}_j$. Since the expectation of a uniform distribution is its midpoint, Equations (14) become

$$\overline{\mathbf{E_t}(.)} = \underline{\mathbf{E_t}(.)} = E_t(.) = \sum_{i,j} mid(\mathbf{u}_i \mathbf{v}_j) \cdot p_{ij} \tag{17}$$

where $mid(.)$ is the midpoint of its interval argument. Then (15) becomes the stronger pair of constraints

$$\begin{aligned} \underline{\mathbf{E_g}} &\leq E_t(.) \\ \overline{\mathbf{E_g}} &\geq E_t(.). \end{aligned} \tag{18}$$

The effect of correlation can be strengthened not only by narrowing $\mathbf{E_t}(.)$, but also by narrowing $\mathbf{E_g}$. A way to narrow $\mathbf{E_g}$ is to accept as inputs point value(s) for expectations and variances $\mu_u = E(u)$, $\mu_v = E(v)$, $\sigma_u^2 = [E(u^2) - E(u)^2]$, and/or $\sigma_v^2 = [E(v^2) - E(v)^2]$, instead of calculating intervals for them from the discretized marginals as in step 1 of Section 3. If these were all point values then the width of $\mathbf{E_g}$ would be controlled by the width of $\boldsymbol{\rho}$, and if ρ was a number then $\mathbf{E_g}$ would be a number (call it E_g) as well.

Since narrowing either $\mathbf{E_t}(.)$ or $\mathbf{E_g}$ tends to strengthen the effects of correlation, a third approach that narrows both is to use a finer discretization for the marginals. Finer discretizations narrow $\mathbf{E_g}$ by narrowing $\mathbf{E}(u), \mathbf{E}(v)$, $\mathbf{E}(u^2)$, and $\mathbf{E}(v^2)$ in Equation (13), and also narrow computations of $\mathbf{E_t}(.)$ by narrowing the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s, resulting in narrower $\mathbf{u}_i \mathbf{v}_j$ terms

$\mathbf{v}_2 = [100, 100]$ $p = 0.5$	$u + v = [101, 101]$ $p_{12} =?$	$u + v = [200, 200]$ $p_{22} =?$
$\mathbf{v}_1 = [1, 1]$ $p = 0.5$	$u + v = [2, 2]$ $p_{11} =?$	$u + v = [101, 101]$ $p_{21} =?$
$u + v \nearrow$	$\mathbf{u}_1 = [1, 1]$ $p = 0.5$	$\mathbf{u}_2 = [100, 100]$ $p = 0.5$

Constraint name	**Equation**
Top row	$p_{12} + p_{22} = 0.5$
2nd row	$p_{11} + p_{21} = 0.5$
2nd column	$p_{12} + p_{11} = 0.5$
Right column	$p_{22} + p_{21} = 0.5$

Table 5: (*top*) joint distribution tableau for a simple problem. (*Bottom*) the linear constraints implied by the tableau.

in Equations (14). Other ways of expressing partial information about dependency, including identification of useful assumptions besides correlation, and when those assumptions are reasonable to make, are likely to enable additional progress in narrowing envelopes around derived distributions.

4 Examples

We start with an example that is simple enough to go through in full detail, followed by another example of more realistic complexity.

4.1 A Basic, Detailed Example

Let the distribution for value u consist of two impulses of equal probability: $\mathbf{u}_1 = [1, 1]$ and $\mathbf{u}_2 = [100, 100]$, with $p(u \in \mathbf{u}_1) = p(u \in \mathbf{u}_2) = 0.5$, and let the distribution describing v be the same as for u. The joint distribution tableau is shown in Table 5. First the envelopes for the case of unknown dependency are derived. Then correlation is added as a constraint and we show how this reduces the separation between the envelopes.

4.1.1 Unknown dependency condition

The left envelope may be derived as follows.

- For $u + v < 2$, $\overline{\mathbf{F}_{u+v}}(.) = 0$ because $u + v$ cannot be below 2.
- For $u + v \in [2, 101)$, only p_{11} contributes its probability to $\overline{\mathbf{F}_{u+v}}(.)$, and its maximum possible value is 0.5. This is because $p_{11} = 0.5$ is consistent with the row and column constraints, shown in Table 5, by setting $p_{11} = p_{22} = 0.5$ and $p_{12} = p_{21} = 0$, while any value for p_{11} over 0.5 would immediately violate the 2nd row and 2nd column constraints. Thus $\overline{\mathbf{F}_{u+v}}(.) = 0.5$ in this case.
- For $u + v \in [101, 200)$, p_{11}, p_{12}, and p_{21} contribute to $\overline{\mathbf{F}_{u+v}}(.)$. Their sum $p_{11} + p_{12} + p_{21}$ can be as high as 1 while remaining consistent with the row and column constraints, by setting $p_{12} = p_{21} = 0.5$ and $p_{11} = p_{22} = 0$. Thus $\overline{\mathbf{F}_{u+v}}(.) = 1$ in this case.

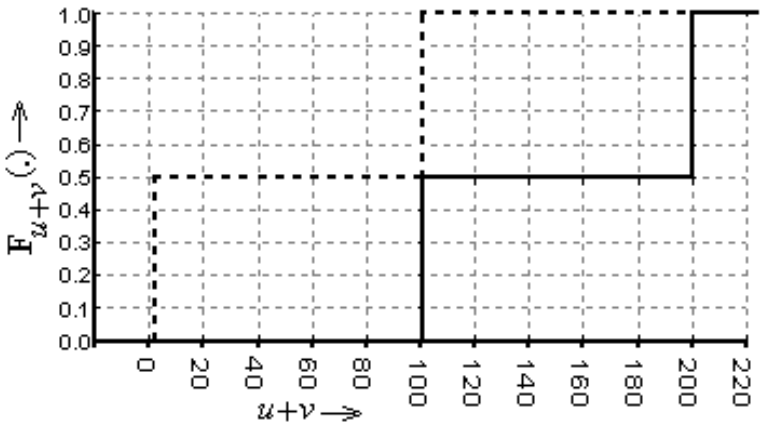

Figure 3: envelopes around the CDF of $u + v$, for the joint distribution tableau of Table 5.

- For $u + v \geq 200$, $\overline{\mathbf{F}_{u+v}}(.) = 1$ because $u + v$ must be at or below 200.

The right envelope may be derived as follows.

- For $u + v < 2$, $\overline{\mathbf{F}_{u+v}}(.) = 0$ because $u + v$ cannot be below 2.
- For $u + v \in [2, 101)$, only p_{11} contributes to $\underline{\mathbf{F}_{u+v}}(.)$. The minimum possible value of p_{11} is 0 because the row and column constraints are all satisfied if we set $p_{11} = p_{22} = 0$ and $p_{12} = p_{21} = 0.5$. Thus $\underline{\mathbf{F}_{u+v}}(.) = 0$ in this case.
- For $u + v \in [101, 200)$, p_{11}, p_{12}, and p_{21} contribute to $\underline{\mathbf{F}_{u+v}}(.)$. Their sum $p_{11} + p_{12} + p_{21}$ can be as low as 0.5 while remaining consistent with the row and column constraints, by setting $p_{12} = p_{21} = 0$ and $p_{11} = p_{22} = 0.5$. Any value below 0.5 for the sum would immediately violate the 2nd row and 2nd column constraints. Thus $\underline{\mathbf{F}_{u+v}}(.) = 0.5$ in this case.
- For $u + v \geq 200$, $\underline{\mathbf{F}_{u+v}}(.) = 1$ because $u + v$ must be at or below 200.

The envelopes are shown in Figure 3. Next we show how correlation narrows the separation of these envelopes.

4.1.2 Effect of correlation

Let us illustrate how correlation works step by step, extending the example just detailed by incorporating the information that $\rho \in [0.7, 1]$. From this, and the joint distribution tableau of Table 5, $\mathbf{E_g}$ may be calculated by substituting intervals into Equation (13) as follows:

$$\mathbf{E_g} = \frac{[1,1]+[100,100]}{2} \cdot \frac{[1,1]+[100,100]}{2}$$
$$+[0.7,1] \cdot \sqrt{\left(\frac{[1,1]^2+[100,100]^2}{2} - \left(\frac{[1,1]+[100,100]}{2}\right)^2\right) \cdot \left(\frac{[1,1]^2+[100,100]^2}{2} - \left(\frac{[1,1]+[100,100]}{2}\right)^2\right)}$$
$$= [4265.425, 5000.5].$$

Next, values are substituted from the interior cells of the joint distribution tableau of Table 5 into Equation (14) to get an expression for $\mathbf{E_t}(.)$, as follows.

$$\begin{aligned}\mathbf{E_t}(.) &= p_{11} \cdot [1,1] \cdot [1,1] + p_{12} \cdot [1,1] \cdot [100,100] + p_{21} \cdot [100,100] \cdot [1,1] + p_{22} \cdot [100,100] \cdot [100,100] \\ &= p_{11} + 100p_{12} + 100p_{21} + 10^4 p_{22}\end{aligned}$$

Thus $\mathbf{E_t}(.)$ is a thin interval in this example. To signify that, we will consider it a number and use the symbol $E_t(.)$ henceforth. The four constraints of Table 5 are augmented with the following two new constraints derived from the computations for $\mathbf{E_g}$ and $E_t(.)$ just shown, and from Equations (15).

$$4265.425 \leq p_{11} + 100p_{12} + 100p_{21} + 10^4 p_{22} \quad (19)$$

$$5000.5 \geq p_{11} + 100p_{12} + 100p_{21} + 10^4 p_{22}. \quad (20)$$

Applying the new constraints. One can now ask how adding Constraints (19)-(20) to the row and column constraints leads to envelopes that are closer together than for the unknown dependency condition.

The new left envelope may be derived as follows.

- For $u + v < 2$ the earlier conclusion, $\overline{\mathbf{F}_{u+v}}(.) = 0$, is unaffected.
- For $u + v \in [2, 101)$ the earlier conclusion, $\overline{\mathbf{F}_{u+v}}(.) = 0.5$, occurs for $p_{11} = p_{22} = 0.5$ and $p_{12} = p_{21} = 0$, is unchanged because those assignments to the p_{ij}'s imply $E_t(.) = 0.5 + 100 \cdot 0 + 100 \cdot 0 + 10^4 \cdot 0.5 = 5000.5$, and 5000.5 is consistent with Constraints (19)-(20).
- For $u + v \in [101, 200)$ the analysis is more involved. The earlier conclusion based on only the row and column constraints was that $\overline{\mathbf{F}_{u+v}}(.) = p_{11} + p_{12} + p_{21} = 1$ and that this could be achieved by setting $p_{12} = p_{21} = 0.5$ and $p_{11} = p_{22} = 0$. For the present scenario of $\rho \in [0.7, 1]$, however, this result is too high because those assignments to the p_{ij}'s lead to the following calculation.

  $$E_t(.) = 1 \cdot 0 + 100 \cdot 0.5 + 100 \cdot 0.5 + 10^4 \cdot 0 = 100 \quad (21)$$

  which violates Constraint (19). The reason is that these assignments to the p_{ij}'s allocate all the probability for value $u + v$ in Figure 5 to p_{12} and p_{21}, which are in the cells for which one marginal has value 1 and the other has value 100. Thus when a value of one marginal is low the value of the other is high. This allocation is inconsistent with the given correlation of [0.7,1] which, being positive, requires u and v to tend to be either both low or both high.

  To calculate a new value of $\overline{\mathbf{F}_{u+v}}(.)$ for $u+v \in [101, 200)$ given $\rho \in [0.7, 1]$, we can derive and solve simultaneous equations on the p_{ij}'s by hand or, as Statool does, invoke linear programming on a computer. For illustration we do it next using simultaneous equations.

  The extreme of assigning all probability to p_{12} and p_{21} and no probability to p_{11} and p_{22}, which gave the envelope height calculated earlier for the unknown dependency condition, is not possible for $\rho \in [0.7, 1]$ as shown by Equation (21). We wish to reduce the sum $p_{11} + p_{12} + p_{21}$ (hence increasing p_{22}) just enough to raise $E_t(.)$ from 100 up to 4265.425, because this will result in the maximum possible assignment to $p_{11} + p_{12} + p_{21}$ that is consistent with $E_t(.) = p_{11} + 100p_{12} + 100p_{21} + 10^4 p_{22} \in [4265.425, 5000.5]$, as required by Constraints (19)-(20). To do this we use, as one of the simultaneous equations, $p_{11} + 100p_{12} + 100p_{21} + 10^4 p_{22} = 4265.425$. Solving this simultaneously with the constraint equations of Table 5 gives $p_{11} + p_{12} + p_{21} = 0.425 + 0.075 + 0.075 = 0.575$.

  The conclusion is that, for $u + v \in [101, 200)$ and $\rho \in [0.7, 1]$, the left envelope height $\overline{\mathbf{F}_{u+v}}(.)$ is 0.575, which is considerably lower than its value of 1 under the unknown dependency condition.
- For $u + v \geq 200$, the earlier conclusion that $\overline{\mathbf{F}_{u+v}}(.) = 1$ is unaffected.

The new right envelope may be derived as follows.

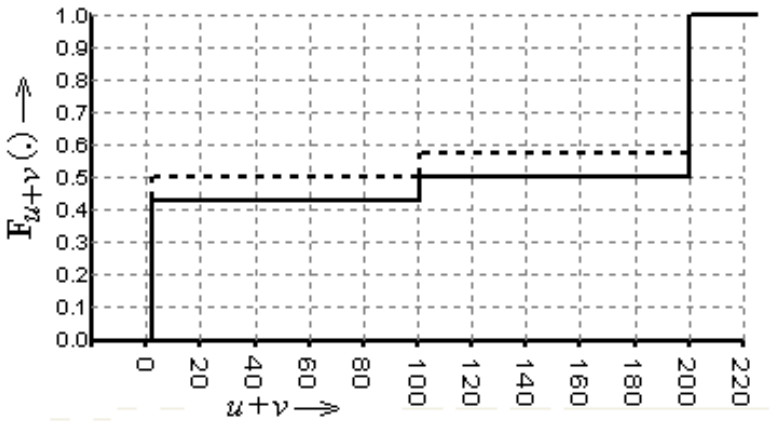

Figure 4: envelopes around the CDF of $u + v$, for the joint distribution tableau of Table 5, given $\rho \in [0.7, 1]$.

- For $u + v < 2$ the earlier conclusion, $\overline{\mathbf{F}_{u+v}} = 0$, is unaffected.
- For $u + v \in [2, 101)$, only p_{11} contributes to $\underline{\mathbf{F}_{u+v}}(.)$. The minimum possible value of 0 for p_{11} found for the unknown dependency condition is too low for $\rho \in [0.7, 1]$. This is because $p_{11} = 0$ implies $p_{22} = p_{11} = 0$ and $p_{12} = p_{21} = 0.5$ due to the constraints shown in Table 5, just as in the discussion of $u + v \in [101, 200)$ for the left envelope, above. There, we moved as small as possible an amount of probability out of $p_{11} + p_{12} + p_{21}$, which was the sum of the contributing cell probabilities. This is the same as moving as small a probability as possible into p_{22}, the only non-contributing cell and thus the complement of the contributing cells. Here, we wish to move as small as possible an amount into p_{11}, not p_{22}, but because the constraints in Table 5 imply $p_{11} = p_{22}$, the resulting allocation of probabilities among the interior cells is actually the same. Thus, as above, Constraints (19)-(20) in conjunction with the constraints of Table 5 imply a minimum value for $p_{11} = p_{22}$ of $1 - (p_{11} + p_{12} + p_{21}) = 1 - 0.575 = 0.425$. Therefore for $u + v \in [2, 101)$, when $\rho \in [0.7, 1]$, $\underline{\mathbf{F}_{u+v}}(.) = 0.425$. This is considerably higher than its value of 0 under the unknown dependency condition.
- For $u + v \in [101, 200)$, the earlier conclusion that $\underline{\mathbf{F}_{u+v}}(.) = 0.5$ occurs for $p_{12} = p_{21} = 0$ and $p_{11} = p_{22} = 0.5$ is unchanged, because those assignments to the p_{ij}'s imply $E_t(.) = 0.5 + 100 \cdot 0 + 100 \cdot 0 + 10^4 \cdot 0.5 = 5000.5$, which is consistent with Constraints (19)-(20).
- For $u + v \geq 200$ the earlier conclusion, $\underline{\mathbf{F}_{u+v}}(.) = 1$, is unaffected.

The envelopes around the CDF of $u + v$ when $\rho \in [0.7, 1]$ are shown in Figure 4. They are closer together than for the unknown dependency condition shown in Figure 3. For ease of exposition the example just described used a joint distribution tableau containing numbers (or strictly speaking, thin intervals). If the marginal intervals are widened, giving weaker specifications for the inputs, wider envelopes around the CDF of $u + v$ result (Figure 5).

4.2 A More Complex Example

Here we show the effects of different correlation conditions using inputs with realistically detailed discretizations. Figures 6 and 7 show two discretized distributions. Let u and v be values drawn from the skewed distribution and the bimodal distribution, respectively. (Bimodal distributions can find application in describing system parameters that are controlled to stay within an

$\mathbf{v}_2 = [99, 101]$ $p = 0.5$	$u + v \in [99, 103]$ $p_{12} = ?$	$u + v \in [198, 202]$ $p_{22} = ?$
$\mathbf{v}_1 = [0, 2]$ $p = 0.5$	$u + v \in [0, 4]$ $p_{11} = ?$	$u + v \in [99, 103]$ $p_{21} = ?$
$v \uparrow \quad u + v \nearrow$ $u \rightarrow$	$\mathbf{u}_1 = [0, 2]$ $p = 0.5$	$\mathbf{u}_2 = [99, 101]$ $p = 0.5$

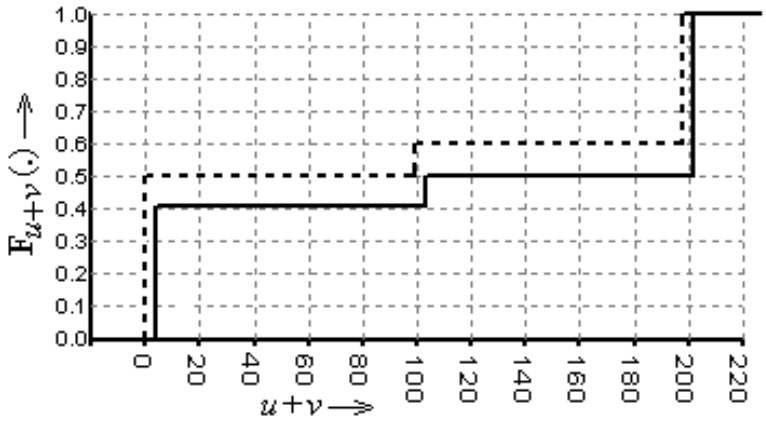

Figure 5: (*top*) joint distribution tableau like that of Table 5 except that the intervals are wider. (*Bottom*) envelopes around the CDF of $u + v$ for the joint distribution tableau at top, assuming $\rho \in [0.7, 1]$. These envelopes are wider than the envelopes in Figure 4 because the $\mathbf{u}_i$'s and $\mathbf{v}_j$'s here specify the PDFs for u and v more weakly, with widths of 2 instead of 0 as in Table 5.

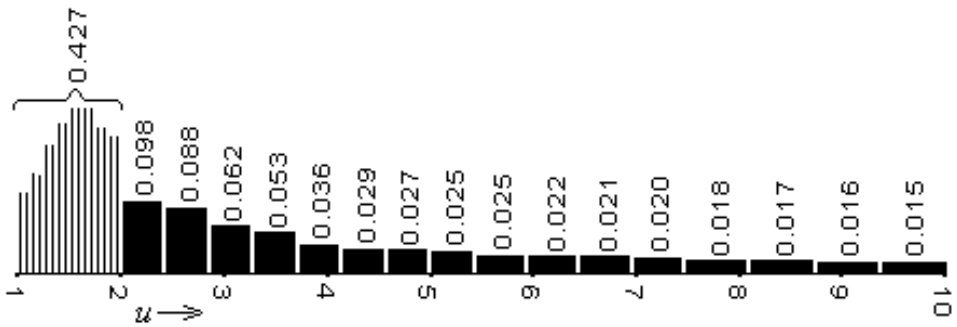

Figure 6: a discretized input distribution. The flat tops of the bars are an artifact of the graphical representation and do not imply uniform (or any other) distribution of probability over the domain of any given bar.

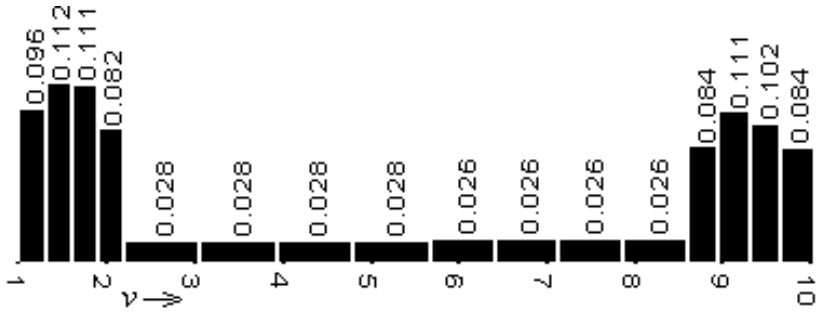

Figure 7: discretization of a bimodal PDF to be used as a divisor.

allowable range. As the parameter wanders within this range, it often approaches the endpoints of the range, activating a control mechanism that prevents it from passing those endpoints. As a result the parameter may tend to spend more time near the endpoints of the range than in the middle.) We further specify that $\mu_u \in [3, 3.1]$, $\mu_v \in [5.15, 5.25]$, $\sigma_u^2 \in [5, 5.1]$, and $\sigma_v^2 \in [11.4, 11.5]$.

Let $z = \frac{u}{v}$. Assuming that values u and v are independent gives envelopes for z that are relatively close together (Figure 8). The relatively small separation between them occurs because the algorithm automatically bounds the effects of discretization as noted in Section 2. Removing the independence assumption leads to envelopes that are much wider apart (Figure 9).

Specifying that the correlation is negative, that is, that $\rho \in [-1, 0)$, results in envelopes that are slightly narrower (Figure 10) than for the unknown correlation condition. Note for example the rounding of the northwest knee of the left envelope relative to the unknown correlation case in Figure 9. This rounding means we can, for example, rule out the possibility that the CDF has value 1 (i.e. certainty) for some values on the horizontal axis, which could potentially be significant for decision-making. Restricting the sign of the correlation appears to usually be a rather weak constraint, since many different dependency relationships can have correlation measures with the same sign.

Stronger correlations can have greater effects. Figure 11 shows 3 pairs of envelopes superposed. Progressing from weaker to stronger restrictions on correlation, the outermost envelopes bound the possible CDFs for z given $\rho \in [-1, -0.5]$. The 2nd envelope from the left and 2nd envelope from the left bound the possible CDFs given $[-1, -0.8]$. The innermost envelopes bound the possible CDFs given the strongest restriction on correlation, $\rho \in [-1, -0.83]$.

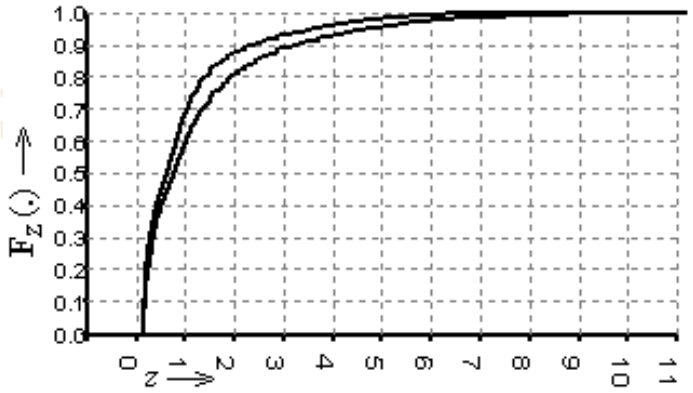

Figure 8: envelopes around the cumulative distribution for z, where $z = u/v$ and u and v are assumed independent. This is a strong assumption that leads to envelopes that are relatively close together.

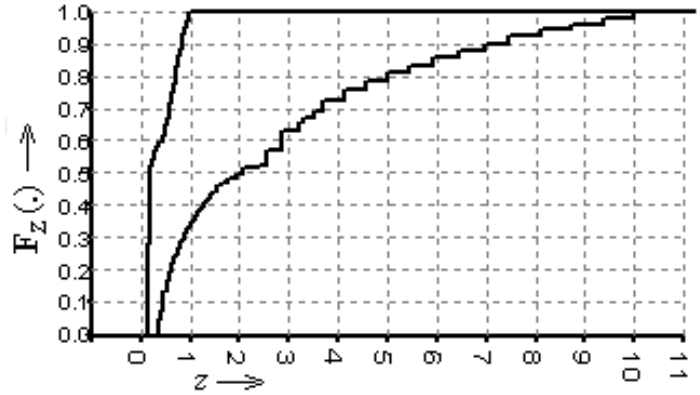

Figure 9: envelopes around the CDF of z, where $z = u/v$ and no assumptions are made about the dependency relationship between u and v. The lack of information about dependency yields envelopes around the cumulative distribution of z that are relatively widely separated.

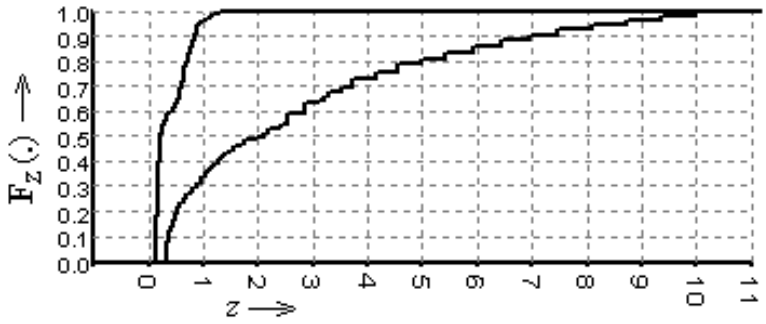

Figure 10: envelopes around the CDF of $z = u/v$, where u and v are assumed to have negative correlation ($\rho < 0$).

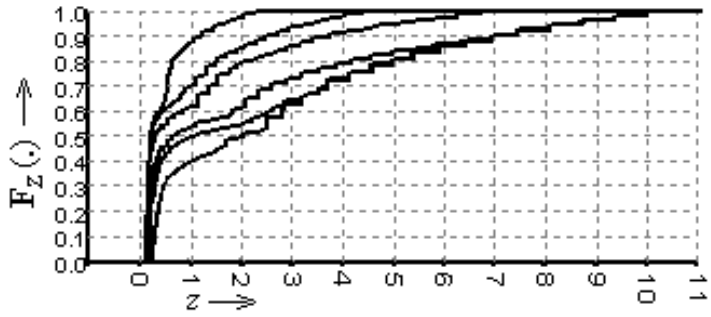

Figure 11: envelopes around the CDF for $z = u/v$ under three correlation conditions. The outermost envelopes are for the weakest of the three, $\rho \in [-1, -0.5]$. The envelopes 2nd from the left and 2nd from the right are for $\rho \in [-1, -0.8]$. The innermost envelopes are for the strongest correlation condition, $\rho \in [-1, -0.83]$.

5 Conclusion

DEnv (Distribution Envelope Determination) is a numerical algorithm for computing envelopes around the space of possible cumulative distribution functions of derived random variables. These are random variables whose values are a function of the values of other random variable(s). Envelopes are appropriate for safely bounding the CDFs of derived random variables when the dependency relationship between the input distributions is not fully known. This is important because often available information is insufficient to reliably justify a particular dependency relationship. Each possible dependency relationship implies some CDF in the family that is bounded by the envelopes. Envelopes can also bound the effects of discretization, which occurs because DEnv requires that input distributions be discretized.

We have previously reported how DEnv can handle the case where the dependency relationship between input distributions is unknown. However, partial information about dependency may be available in the form of values or ranges for correlation. This paper extends the DEnv algorithm to incorporate such information about correlation. Pearson correlation is used because it is the most commonly used kind of correlation. Some examples are provided, showing how correlation can strengthen results relative to those obtained without any information about dependency.

Acknowledgements

The authors thank Gerald Sheblé for discussions regarding needs for advances in the DEnv algorithm and possibilities for its application. Using applications to drive advances in the theory and software for DEnv is an important part of our research strategy. As a result we continue to pursue applications in such areas as competitive bidding [6], financial engineering [17], and electric power generation [4].

The anonymous referees contributed significantly to the exposition. Referee #1 offered comprehensive suggestions for revision which are greatly appreciated. The authors retain full responsibility for any remaining shortcomings. This work has been supported in part by research funding from the Power Systems Engineering Research Center (PSERC).

References

1. Alefeld, G. and J. Herzberger, *Introduction to Interval Computations,* Academic Press, New York, 1983.
2. Berleant, D., Automatically Verified Reasoning with Both Intervals and Probability Density Functions, *Interval Computations* (1993 No. 2), pp. 48-70.
3. Berleant, D. and C. Goodman-Strauss, Bounding the Results of Arithmetic Operations on Random Variables of Unknown Dependency Using Intervals, *Reliable Computing* **4** (2) (1998), pp. 147-165.
4. Berleant, D., J. Zhang, R. Hu, and G. Sheblé, Economic Dispatch: Applying the Interval-Based Distribution Envelope Algorithm to an Electric Power Problem, *SIAM Workshop on Validated Computing 2002 Extended Abstracts*, Toronto, May 23-25, pp. 32-35. http://www.public.iastate.edu/~berleant/.
5. Berleant, D. and J. Zhang, Representation and Problem Solving with the Distribution Envelope Determination (DEnv) Method, *Reliability Engineering and System Safety,* forthcoming. http://www.public.iastate.edu/~berleant/.
6. Cheong, M.-P., Competitive Bidding to Sell Power Under Epistemic Uncertainty About the Competition, master's thesis, Dept. of Electrical and Computer Engineering, Iowa State University, in preparation.
7. Colombo, A.G. and R.J. Jaarsma, A Powerful Numerical Method to Combine Random Variables, *IEEE Transactions on Reliability* **R-29** (2) (June 1980), pp. 126-129.
8. Ferson, S., *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, Lewis Press, Boca Raton, 2002.
9. Ferson, S., What Monte Carlo Methods Cannot Do, *Journal of Human and Ecological Risk Assessment* **2** (4) (1996), pp. 990-1007.
10. Ferson, S. and M. Burgman, Correlations, Dependency Bounds and Extinction Risks, *Biological Conservation* **73** (1995), pp. 101-105.
11. Frank, M.J., R.B. Nelsen, and B. Schweizer, Best-Possible Bounds for the Distribution of a Sum – a Problem of Kolmogorov, *Probability Theory and Related Fields* **74** (1987), pp. 199-211.
12. Ingram, G.E., E.L. Welker, and C.R. Herrmann, Designing for Reliability Based on Probabilistic Modeling Using Remote Access Computer Systems, *Proceedings 7th Reliability and Maintainability Conference,* American Society of Mechanical Engineers, 1968, pp. 492-500.
13. Moore, R., Risk Analysis Without Monte Carlo Methods, Freiburger Intervall-Berichte, 84/1, 1984, pp. 1-48.
14. Nelsen, R.B., *An Introduction to Copulas,* Lecture Notes in Statistics, Vol. 139, Springer-Verlag, Heidelberg, 1999.
15. Neumaier, A., Clouds, Fuzzy Sets, and Probability Intervals, submitted. www.mat.univie.ac.at/ neum/papers.html.
16. Red-Horse, J. and A.S. Benjamin, A Probabilistic Approach to Uncertainty Quantification with Limited Information, *Reliability Engineering and System Safety,* forthcoming.
17. Sheblé, G. and D. Berleant, Bounding the Composite Value at Risk for Energy Service Company Operation with DEnv, an Interval-Based Algorithm, *SIAM Workshop on Validated Computing 2002 Extended Abstracts*, Toronto, May 23-25, pp. 166-171. http://www.public.iastate.edu/~berleant/.
18. Springer, M.D., *The Algebra of Random Variables,* John Wiley and Sons, New York, 1979.

19. Statool software.
class.ee.iastate.edu/berleant/home/Research/Pdfs/versions/statool/distribution/index.htm.

20. Tajar, A., M. Denuit, and P. Lambert, Copula-Type Representations for Random Couples with Bernoulli Margins, Discussion Paper 0118, Institut de Statistique, Université Catholique de Luvain, 2001. www.stat.ucl.ac.be/ISpublications.html.

21. Williamson, R.C. and T. Downs, Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds, *International Journal of Approximate Reasoning* **4** (1990), pp. 89-158.

22. Wood, A.J. and B.F. Wollenberg, *Power Generation, Operations, and Control, 2nd ed.,* Wiley, Hoboken, 1996.