A Simple Closed Form Estimation of the Cumulative Distribution Function of a Monotone Function of Random Variables

K. David Jamison[1], Weldon A. Lodwick[2] and Michael Kawai[2]
1. Watson Wyatt & Company, 950 17th Street,Suite 1400, Denver, CO 80202, U.S.A.
2. Department of Mathematics, Campus Box 170, University of Colorado, P.O. Box 173364, Denver, CO 80217-3364, U.S.A.
e-mail: Ken_Jamison@WatsonWyatt.com, Weldon.Lodwick@cudenver.edu

Abstract: A simple method for estimating the cumulative distribution function of a monotone function of random variables is presented. The method creates a sequence of bounds on the distribution function that will converge to the distribution function in the limit. Moreover, an approximation computed from and contained in the enclosure is, in many cases, sufficiently close to the solution to stop the process at an early stage. An examples is given to illustrate the method.

Keywords: Probability Theory, Interval Analysis

1 Introduction

We will describe a method for estimating the cumulative distribution function (c.d.f.) for any monotone function of a finite set of random variables. An illustration of the method is presented. When the number of random variables is small (less than 10), the method gives good results in reasonably short order so that we believe it warrants further study. The estimate is in a closed form that might prove useful in stochastic and possibilistic programing problems. We note the similarity between our method with that of R.E. Moore ([6]) and several others after Moore (e.g. [5], [2] and [7]). The method is an application of the ideas discussed in [4].

2 Estimating the c.d.f.

Let $Y = f(X)$ where the $X = (X_1, ..., X_n)$ is a vector of continuous random variables with joint distribution function $F_X(x)$ (i.e.$F_X(x) = prob(X_1 \leq x_1, ..., X_n \leq x_n)$

with marginals F_{X_i} and $G_X(x) = prob(x_1 \leq X_1, ..., x_n \leq X_n)$ (see; for example, [3]). We assume f is continuous and nondecreasing in each x_i (it is a simple adjustment to consider functions that increase in some variables and decrease in others) and that the support of each X_i is bounded with $\sup p(X_i) = \left[F_{X_i}^{-1}(0), F_{Xi}^{-1}(1)\right]$ (note the abuse of notation for convenience).

We construct bounds (upper/lower) and an approximation between the bounds for the c.d.f. of Y. There are three steps to the method. The first step consists of partioning the domain space into smaller blocks. In the second step we construct a bound and an intermediate estimate for the conditional c.d.f of Y for each block of the partition given X is in that block. The final step is to combine the conditional c.d.f.s into the final estimate.

The first step is to construct a partition on the domain, $\left[F_{X_1}^{-1}(0), F_{X_1}^{-1}(1)\right] \times ... \times \left[F_{X_n}^{-1}(0), F_{X_n}^{-1}(1)\right]$. We do this by dividing each interval into subintervals equally spaced in probability (relative to the marginal distributions). For example, to divide $\left[F_{X_i}^{-1}(0), F_{Xi}^{-1}(1)\right]$ into three pieces of equal probability we use $\left[F_{X_i}^{-1}(0), F_{X_i}^{-1}\left(\frac{1}{3}\right)\right]$, $\left[F_{X_i}^{-1}\left(\frac{1}{3}\right), F_{X_i}^{-1}\left(\frac{2}{3}\right)\right]$ and $\left[F_{X_i}^{-1}\left(\frac{2}{3}\right), F_{X_i}^{-1}(1)\right]$ (note we are not concerned with overlap since we have assumed the distribution of X is continuous). The primary consideration in this process is how it effects the size of the problem. If there are n random variables and each variable X_i is divided into k_i subintervals then we will have $\prod_{i=1}^{n} k_i$ conditional c.d.f.s to compute. Therefore it is desirable to minimize the number of subdivisions and only subdivide the variables that influence the results the most.

The second step is to construct the bounds and the estimated conditional c.d.f. for each block of the partition. Let $[b_1, c_1] \times ... \times [b_n, c_n]$ be one such block and let A denote the event X falls in this block. Consider the family of n-dimensional blocks $\left\{\left[b_1, F_{X_1|A}^{-1}(\beta)\right] \times ... \times \left[b_n, F_{X_n|A}^{-1}(\beta)\right] \mid \beta \in [0,1]\right\}$. From our assumption that f is continuous and increasing in each X_i we know that

$$f\left(\left[b_1, F_{X_1|A}^{-1}(\beta)\right] \times ... \times \left[b_n, F_{X_n|A}^{-1}(\beta)\right]\right) = \left[f(b_1, ...b_n), f\left(F_{X_1|A}^{-1}(\beta), ..., F_{X_n|A}^{-1}(\beta)\right)\right]$$

Thus

$$F_{Y|A}\left(f\left(F_{X_1|A}^{-1}(\beta), ..., F_{X_n|A}^{-1}(\beta)\right)\right) \geq F_{X|A}\left(F_{X_1|A}^{-1}(\beta), ..., F_{X_n|A}^{-1}(\beta)\right).$$

Now consider the n-dimensinal block $\left[F_{X_1|A}^{-1}(\beta), c_1\right] \times ... \times \left[F_{X_n|A}^{-1}(\beta), c_n\right]$.
As before, we know that

$$f\left(\left[F_{X_1|A}^{-1}(\beta), c_1\right] \times ... \times \left[F_{X_n|A}^{-1}(\beta), c_n\right]\right) = \left[f\left(F_{X_1|A}^{-1}(\beta), ..., F_{X_n|A}^{-1}(\beta)\right), f(c_1, ...c_n)\right].$$

This gives the inequality

$$1 - F_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right) \geq G_{X|A}\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right).$$

Put together we have

$$F_{X|A}\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right) \leq F_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right) \leq 1 - G_{X|A}\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right).$$

When the random variables, X, are independent this becomes

$$\beta^n \leq F_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right) \leq 1 - (1-\beta)^n.$$

This is a wide envelope, particularly for a large number of variables (large n). We wish to produce a reasonable estimate of the c.d.f. without having to perform the calculations needed reduce the envelope to reasonable width. To do this we select an intermediate value for $F_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right)$ that falls between the upper and lower estimate above. One estimate would be to average these probabilities, i.e. set $F_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right) = \frac{1}{2}\left(F_{X|A}\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right) + 1 - G_{X|A}\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right)$. When the random variables, X, are independent a reasonable intermediate estimate is simply to use β. This works since $\beta^n \leq \beta \leq 1 - (1-\beta)^n$ and has the property that it does not increase the maximum possible error in making a choice of intermediate value. This is so because the maximum of the difference $1 - (1-\beta)^n - \beta^n$ occurs when $\beta = .5$ and at this value the midpoint estimate is $\frac{1}{2}(.5^n + 1 - .5^n) = .5$. So for independent X we use

$$F^{-}_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right) = \beta^n \tag{1}$$

$$\hat{F}_{Y|A}\left(f\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right)\right) = \beta \tag{2}$$

and

$$F^{+}_{Y|A}\left(F^{-1}_{X_1|A}(\beta), ..., F^{-1}_{X_n|A}(\beta)\right) = 1 - (1-\beta)^n \tag{3}$$

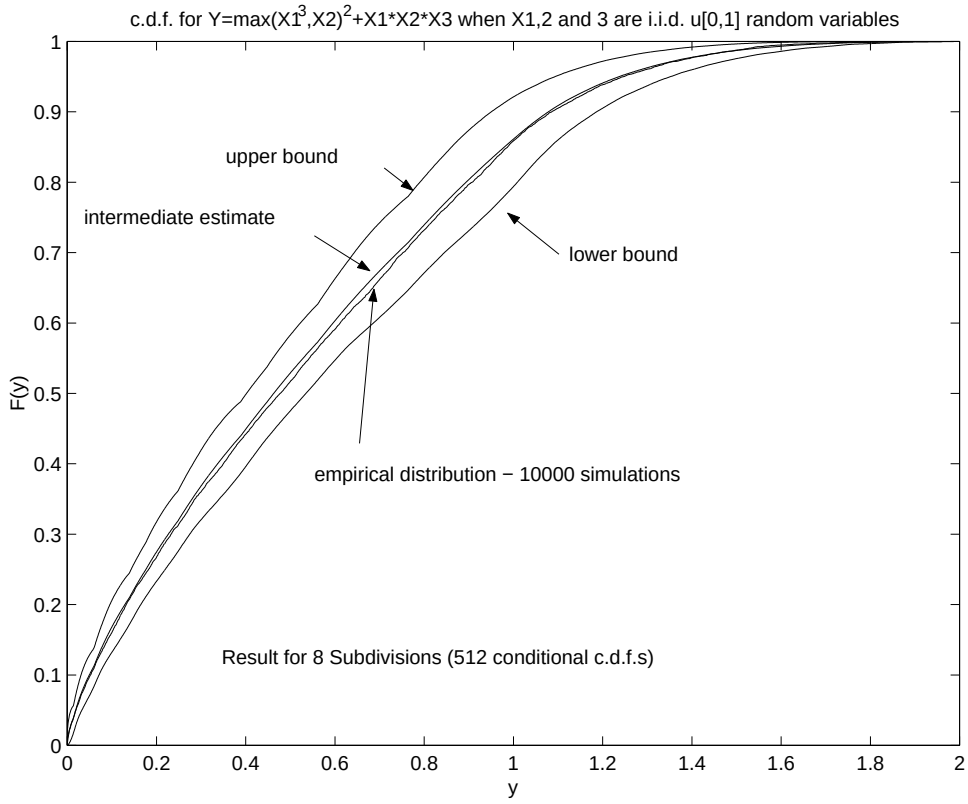
to obtain a lower bound, intermediate estimate and upper bound on the actual c.d.f., $F_{Y|A}(y)$. These are the values that will be used for the rest of this paper. Further refinements are a subject of future research.

The last step of the method is to combine the estimated conditional c.d.f.s in to an estimate of the c.d.f. for Y. Assume we have divided the support of each X_i into subintervals creating a partition of the support of X. If $\{A_j \mid j = 1, m\}$ is the partition (so each A_j is also an n-dimensional block) and if $F^-_{Y|A_i}(y)$, $\hat{F}_{Y|A_i}(y)$ and $F^+_{Y|A_i}(y)$ are the values calculated as above to bound and estimate the c.d.f. for the variable $Y \mid X \in A_i$ then we can combine these c.d.f.s to produce the bounds and estimate for the c.d.f. of interest by , for example, setting $\hat{F}_Y(y) = \sum_{j=1}^m \hat{F}_{Y|A_i}(y) P(X \in A_j)$ (were $P(E)$ equals the probability of the event E). The upper and lower bounds ($F^-_Y(y)$ and $F^+_Y(y)$) are similarly calculated.

It is clear that this process will converge to the actual c.d.f. if the supports for each X_i are subdivided into finer and finer subintervals since $F^-_Y(y)$ and $F^+_Y(y)$ are simply inner and outer measures (relative to the measure defined by F_X) for the region $\{x \mid f(x) \leq y\}$ (see any introductory measure theory text; for example, [1]).

3 Example

Consider $Y = (\max\{X_1^3, X_2\})^2 + X_1X_2X_3$ where X_i are i.i.d. u$[0,1]$ random variables. Then on $A = [a,b] \subseteq [0,1]$ $F_{X|A}(x) = \frac{x-a}{b-a}$ and the inverse of this is $F^{-1}_{X|A}(\beta) = \beta(b-a) + a$. Assume we have subdivided each $[0,1]^3$ into 512 blocks by dividing each interval $[0,1]$ into eight intervals $\left[0, \frac{1}{8}\right]$, $\left[\frac{1}{8}, \frac{2}{8}\right]$,...,$\left[\frac{7}{8}, 1\right]$. Then, for example, on $A = \left[0, \frac{1}{8}\right] \times \left[\frac{7}{8}, 1\right] \times \left[\frac{2}{8}, \frac{3}{8}\right]$ we have $F^{-1}_{X_1|A}(\beta) = \frac{1}{8}\beta$, $F^{-1}_{X_2|A}(\beta) = \frac{1}{8}\beta + \frac{7}{8}$ and $F^{-1}_{X_3|A}(\beta) = \frac{1}{8}\beta + \frac{2}{8}$. Then we bound and estimate the conditional c.d.f. at $y = (\max\{\left(\frac{1}{8}\beta\right)^3, \frac{1}{8}\beta + \frac{7}{8}\})^2 + \frac{1}{8}\beta\left(\frac{1}{8}\beta + \frac{7}{8}\right)\left(\frac{1}{8}\beta + \frac{7}{8}\right)$ by $F^-_{Y|A}(y) = \beta^3$, $\hat{F}_{Y|A}(y) = \beta$ and $F^+_{Y|A}(y) = 1 - (1-\beta)^3$. Then the c.d.f. for Y is estimated by summing over the 512 conditional c.d.f.s multiplied by $\left(\frac{1}{8}\right)^3$, the probability that all three random variables fall into any one of the 512 blocks. The result of this calculation compared to the empirical distribution function for a 10000 simulation is as follows.



4 Conclusion

For a small number of variables, the method gives a reasonable estimate of the c.d.f. and a tight bound on a laptop in several minutes. The intermediate estimate for the c.d.f. for up to 14 independent random variables has been calculated to a good degree of accuracy in several minutes on a laptop even while the upper and lower bounds are still quite wide. Further research is needed to make the method of greater use when a large number of variables are involved, to extend the results to non-monotone functions and to the selection of better intermediate estimates.

References

[1] S.K. Berberian, *Measure and Integration* (Chelsea Publishing Company, Bronx, New York, 1970).

[2] D. Berleant, Automatically Verified Reasoning with Both Intervals and Probability Density Functions, *Interval Computation* 2 (1993), pp. 48-70.

[3] L. Breiman, *Probability* (S.I.A.M., 1992).

[4] K.D. Jamison and W.A. Lodwick, The Construction of Consistent Possibility and Necessity Measures, *Fuzzy Sets and Systems (accepted 2002)*.

[5] A.S. Moore, *Interval Risk Analysis of Real Estate Investment: A Non-Monte Carlo Approach*

[6] R.E. Moore, *Risk Analysis without Monte Carlo methods,* Freiburger Intervall-Berichte 1 (1984) 1-48.

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