Departmental Technical Reports (CS)   Department of Computer Science

1-1-2006

Modeling Correlation and Dependence Among Intervals

Scott Ferson

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Modeling Correlation and Dependence Among Intervals

Scott Ferson and Vladik Kreinovich
Applied Biomathematics, Setauket, New York 11733 USA, scott@ramas.com
University of Texas, El Paso, Texas 79968 USA, vladik@utep.edu

Abstract: This note introduces the notion of dependence among intervals to account for observed or theoretical constraints on the relationships among uncertain inputs in mathematical calculations. We define dependence as any restriction on the possible pairings of values within respective intervals and define nondependence as the degenerate case of no restrictions (which we carefully distinguish from independence in probability theory). Traditional interval calculations assume nondependence, but alternative assumptions are possible, including several which might be practical in engineering settings that would lead to tighter enclosures on arithmetic functions of intervals. We give best possible formulas for addition of intervals under several of these dependencies. We also suggest some potentially useful models of correlation, which are single-parameter families of dependencies, often ranging from the identity dependence ($u=v$) representing maximal correlation, through nondependence, to opposite dependence ($1-u=v$) representing maximally negative correlation.

1. Introduction

Interval analysis has an inadequate model of dependence between variables. Because of this deficiency, many analysts discount the utility of interval arithmetic in propagating uncertainty through mathematical expressions because it does not account for natural dependencies that can occur between input values. Many reject interval methods and appeal instead to probability theory because it provides a well developed model of dependence in terms of correlations and the general theory of copulas (Nelsen 1999). This perceived advantage of probabilistic over interval methods is undeserved, however, because interval analysis *could* also offer a model of dependence, and it would be considerably simpler and perhaps more workable than that required for event probabilities or random numbers.

There are two uses of a model of dependence among intervals. The first is to account for dependencies that exist between distinct inputs. Such dependencies can be implied by the physical or biological mechanisms governing the underlying system. For instance, if both the size and mass of a component are interval inputs in a calculation, it is likely there is a connection between these two inputs such that large values of one are associated with large values of the

other and that precludes certain contrary combinations of values within their intervals. For other variables there might be reasons why large values of one cannot co-occur with large values of another. Dependencies such as these might be deduced from the mathematical relationships between the variables. They might alternatively be evidenced by empirical information, or simply asserted a priori by the analyst. In any case, it is legitimate and essential to take account of these dependencies if doing so tightens the interval outputs of analysis.

Although not a primary focus of this note, the second use of a model of dependence among intervals is as underpinning for a strategy to address the repeated parameter issue (also known as the "dependence" issue) in which a single interval input appears multiple times within a mathematical expression. For example, the terms in the expression $A - A^2$ are dependent in that knowing A's value tells us the value of A^2 exactly. Such dependencies arise because of mathematical identities or repeated variables in expressions, rather than empirical dependencies discussed above. One could argue that one kind of dependence is a special case of the other kind of dependence, and they are clearly closely intertwined.

2. Dependence between intervals

So what is dependence between uncertain numbers characterized by intervals? We define dependence as any restriction on the possible pairings of the uncertain numbers. An interval *dependence relation D* is a subset of the unit square $U = [0,1] \times [0,1] = \{(u,v) : u \in [0,1], v \in [0,1]\}$ such that there exists in the relation at least one pair (u,v) for every value of u and v. That is, $D \subseteq U$ is a dependence relation if and only if, for any $u \in [0,1]$, there exists some pair $(u,v) \in D$ for some $v \in [0,1]$, and, likewise, for any $v \in [0,1]$, there is a pair $(u,v) \in D$ for some $u \in [0,1]$. Consider two intervals $A = [a_1, a_2]$ and $B = [b_1, b_2]$. We say that A and B are dependent according to a dependence relation D if

$$f(A, B) = \{c : c = f(a, b), \text{ where } a = u\,(a_2 - a_1) + a_1, b = v\,(b_2 - b_1) + b_1, \text{ and } (u,v) \in D\}$$

for all binary functions f. In this case, A and B are said to have the dependence D. Any pair of values (a,b) is called a *possible pair* from the intervals A and B if $a \in A$, $b \in B$, and $((a - a_1)/(a_2 - a_1), (b - b_1)/(b_2 - b_1)) \in D$.

We use $\mathcal{D}$ to denote the set of all such dependence relations, of which $U \in \mathcal{D}$ is a privileged special case. If a dependence relation is all of U, it is called the *noninteractive dependence relation* or, more simply, the *all-pairs relation*. It is the largest possible dependence in that it encloses all other possible dependence relations. We can say that intervals having this degenerate

relation are *nondependent*. (We conscientiously refrain from calling such intervals 'independent' because this term already has a firmly entrenched meaning in probability theory that is not equivalent to—and indeed is quite different from—nondependence.)

If there is only one pair in the set for each value of u and only one pair for each v, it is called a *one-pair dependence relation*. There are two special cases of one-pair relations that are especially important. The first is the identity relation P = $\{(u,v) : u = v, u \in [0,1], v \in [0,1]\}$. This is the case of perfect dependence between the two intervals. Low values of one interval are perfectly paired with low values of the other, and high values of one are paired with high values of the other. The second special case of a one-pair relation is the opposite relation O = $\{(u,v) : 1-u = v, u \in [0,1], v \in [0,1]\}$ which reverses the association so high values of one variable are paired with low values of the other. Both of these special cases are monotone relations, but not all one-pair relations are so well behaved. Even if the value within A perfectly determines the associated value within B and vice versa, their dependence may still be very complicated. The notion of "shuffles" (Nelsen 1999) from probability theory generalizes to interval dependence.

Between the degenerate all-pairs dependence relation and various possible one-pair dependence relations there is a huge variety of dependence relations. Indeed, this variety is infinite-dimensional, although it is vastly less complex than the analogous diversity in copulas modeling dependence between random numbers in probability theory. The key to developing practical strategies for handling dependence among intervals is to define classes or families of dependence that are appropriate models of the kinds of associations commonly encountered. The next section introduces some candidates.

3. Correlation models

A *complete* model of correlation is any map ρ from $[-1,+1]$ to $\mathcal{D}$ (the set of all bivariate dependence relations) such that $\rho(-1) = O$, $\rho(0) = U$, and $\rho(1) = P$. There are infinitely many such maps (just as there are in the analogous probability theory). Nevertheless, it is useful to identify some models of correlation that might be workable in practical engineering settings. For instance, it might be convenient to define the family of dependence relations depicted in *Figure 1*. The figure shows eleven dependence relations, ranging from O at the far left to P at the far right. Each dependence relation is depicted as an area in black within the unit square. The abscissas are the u values and the ordinates are the v values.

$r = -1$ $r = 0$ $r = +1$
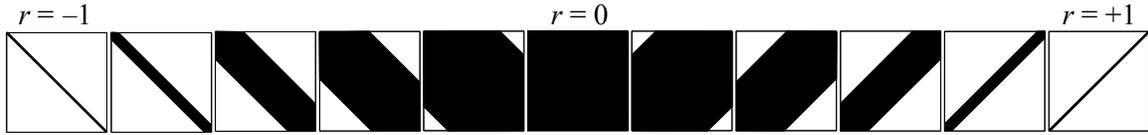

Figure 1. A complete model of correlation between intervals.

In this family, there is a dependence relation for each value of the correlation r from negative one through zero and on to positive one. In this case, the dependence relation for a given r is defined as

$$D(r) = \{(u,v) : \max(0, -u-r, u-1+r) \le v \le \min(1, u+1-r, -u+2+r), u \in [0,1], v \in [0,1]\}.$$

The signal characteristic of the model of correlation represented by this parameterized family of dependence relations is the way in which pairs are excluded that would contradict the assertion of correlation at magnitude r: the counterindicated corners of the dependence relation are shaved away.

There are actually many complete models of correlation that are possible. For example, *Figure 2* shows four different families, each of which smoothly morph from the opposite dependence O for a correlation r of -1 though the all-pairs dependence at correlation zero to the perfect dependence P at correlation $+1$. These families are composed of relations having rhomboidal shapes with straight-line edges. Other families could be devised out of other curved shapes as well.
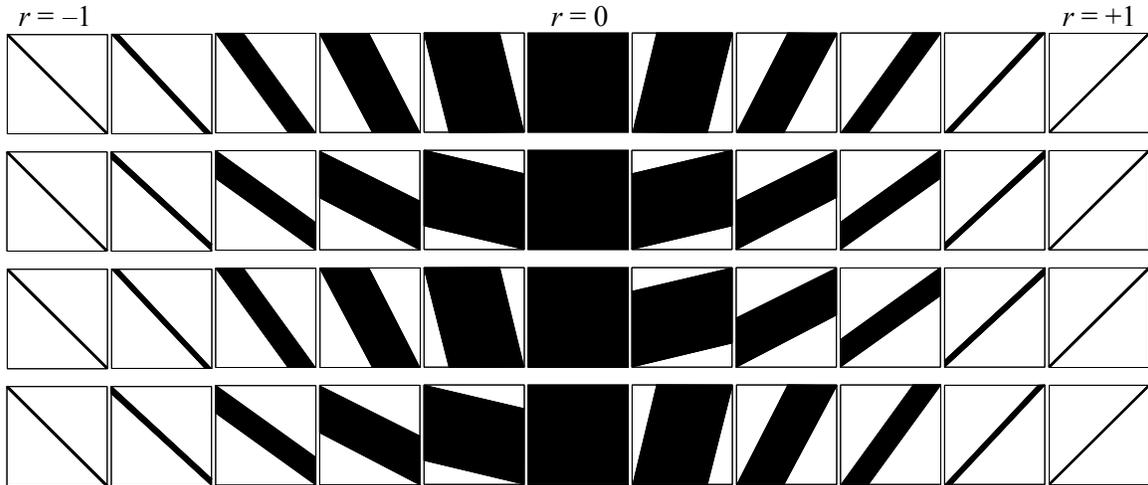
$r = -1$ $r = 0$ $r = +1$

Figure 2. Four alternative complete models of correlation.

A *non-complete* model of correlation would be a map from a proper subset of $[-1,+1]$ to $\mathcal{D}$, or a map from $[-1,+1]$ to $\mathcal{D}$ that didn't send -1, 0 and $+1$ to O, U and P respectively. One non-complete model of correlation that will likely be very useful in practical problems is the ellipse model (Chernousko 1988, 1994; Kreinovich et al. 2005, 2006). This model maps all of the range $[-1,+1]$ and it goes from O and P, but its dependence for zero correlation is not U. Instead, it is the inscribed circle $E_0 = \{(u,v) : (u-\frac{1}{2})^2 + (v-\frac{1}{2})^2 \le \frac{1}{4}\}$. As the correlation coefficient varies from zero to $+1$, the dependence relation is a rotated ellipse inscribed within U. In the limit, as the correlation reaches $+1$, it becomes a degenerate rotated ellipse equivalent to the perfect dependence relation P. Likewise the negative correlations go from the circle to the opposite dependence O. This family of ellipses is depicted in *Figure 3*. It is parameterized by the point u^* of the ellipse's tangency with the *u*-axis (where $v = 0$). Because this point ranges over $[0,1]$, we can define another correlation index $r = 1 - 2u^*$, ranging over $[-1,+1]$.
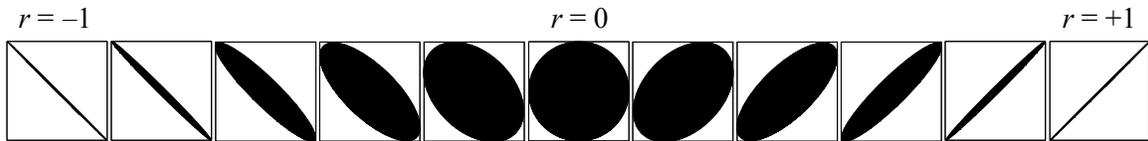

Figure 3. Elliptic family of dependence relations.

Given an elliptic correlation r, the dependence relation is the interior of the ellipse $E(r) \subseteq U$, which is tangent to the *u*-axis at $u^* = (1-r)/2$. This dependence relation is

$$E(r) = \{(u,v) : 4\big((u+v-1)^2 - 2(1+r)(u-\tfrac{1}{2})(v-\tfrac{1}{2})\big)/(1-r^2) \le 1,\ u \in [0,1],\ v \in [0,1]\}.$$

Chernousko (1988, 1994) and Kreinovich et al. (2005, 2006) considered such ellipses for modeling dependence among intervals. Kreinovich et al. (2006) reviewed the use of an elliptic model of interval dependence in quadratic response surface models.

There are many, many other dependence families that might be useful. When, for example, an interval expression involves repeated subexpressions inducing a mathematical dependence, the relevant family of dependence relations represents the mathematical relationship. Consider, for instance, intervals A and A^2. Depending on the numerical values within A, their dependence must be an arc of a parabola and might be one of the dependence relations depicted in *Figure 4*.
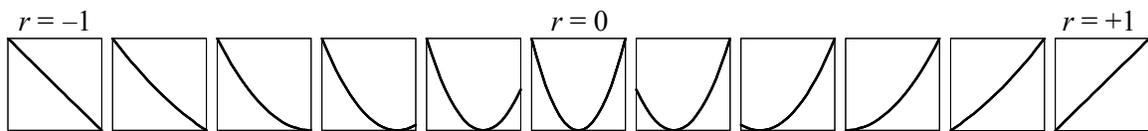

Figure 4. Parabolic family of dependence relations.

These dependencies can be called the parabolic family of dependence relations. A parameterization of the family is
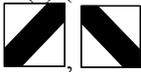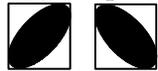
$$Q(r) = \{(u,v) : ((u - \lambda)^2 - q) / (\max(\lambda^2, (1 - \lambda)^2) - q) = v, u \in [0,1], v \in [0,1]\}.$$

where $\lambda = \tan(-\pi r/2) + \frac{1}{2}$ is the location on the u-axis of the parabola's minimum, and q is zero if $0 \leq \lambda \leq 1$, or $\min(\lambda^2, (1-\lambda)^2)$ otherwise. Some of these dependencies are one-pair relations (when they represent only one branch of the parabola), in which case calculations may be relatively easy, but this is not always so. Because the dependence relation is scaled on the unit square, this family of dependences can be parameterized by a single-dimensional scalar value that depends on whether the interval A straddles zero or not.

In principle, other intervals could have parabolic dependence as well. For instance, the interval $B = [4,11]$ could not be a square of the interval $A = [3,5]$ because their ranges would be inconsistent, but these two intervals could have a parabolic dependence if the pairings of $a \in A$ and $b \in B$ were constrained so that $((a - 3)/2, (b - 4)/7) \in Q(r)$ for some r as depicted in Figure 4.

4. Arithmetic operations under specified dependence

Accounting for the dependence between intervals can improve the enclosures that can be computed for arithmetic expressions that involve them, and the numerical results can be considerably tighter than would be obtained by applying the default methods of interval arithmetic that do not consider dependence. The table below gives formulas for the sum of $A = [a_1, a_2]$ and $B = [b_1, b_2]$ under a variety of dependence relations between them. On the left side of the table are given the name of the dependence relation, a graphical depiction of its shape and the constraints that define it (in terms of u and v, which are each implicitly assumed to lie within $[0,1]$). On the right side of the table are formulas to compute best-possible bounds on the sum $A+B$. Some of the formulas involve the envelope function $\text{env}(x,y) = [\min(x, y), \max(x, y)]$, and the proportional component function $\text{w}([x_1,x_2], p)$ which is $p(x_2 - x_1) + x_1$, or just x_1 if p is less than zero, or x_2 if p is greater than one.

Dependence	Addition formula
P (perfect) $u = v$	$[a_1+b_1, a_2+b_2]$
O (opposite) $1 - u = v$	$\text{env}(a_1+b_2, a_2+b_1)$
D(r) (correlated) , $\max(-u-r,u-1+r)\leq v\leq$ $\min(u+1-r,-u+2+r)$	$[\text{env}(\text{w}(A, -r)+b_1, a_1+\text{w}(B,-r)), \text{env}(a_2+\text{w}(B,1+r),\text{w}(A,1+r)+b_2)]$
E(r) (elliptic) , $4((u+v-1)^2-2(1+r)$ $(u-\frac{1}{2})(v-\frac{1}{2}))/(1-r^2)\leq 1$	$\text{env}(p-q^-, p-q^+, -p+q^+, -p+q^-)+(x_1+x_2+y_1+y_2)/2,$ where $p = \sqrt{4z/(((y^2-4xz)/(y-2z))^2-y^2+4xz)}, \quad q^\pm = yp\pm\sqrt{y^2p^2-4z(xp^2-1)})/2z,$ $x = 4/(a_2-a_1)^2(1-r^2), \quad y = -8/(a_2-a_1)(b_2-b_1)(1-r^2), \quad z = 4/(b_2-b_1)^2(1-r^2)$
Upper, left $u \leq v$	$[a_1+b_1, a_2+b_2]$
Lower, left $1 - u \geq v$	$\text{env}(a_2+b_1, \text{env}(a_1+b_2, a_1+b_1))$
Upper, right $1 - u \leq v$	$\text{env}(a_2+b_1, \text{env}(a_1+b_2, a_2+b_2))$
Lower, right $u \geq v$	$[a_1+b_1, a_2+b_2]$
Diamond $\lvert u - \frac{1}{2}\rvert + \lvert v - \frac{1}{2}\rvert \leq \frac{1}{2}$	$[\text{env}(a_1+\text{w}(B,\frac{1}{2}), \text{w}(A,\frac{1}{2})+b_1), \text{env}(a_2+\text{w}(B,\frac{1}{2}), \text{w}(A,\frac{1}{2})+b_2)]$

U (nondependent)	
⬛	$[a_1+b_1, a_2+b_2]$
(u,v)	

This and comparable tables for other arithmetic operations such as subtraction, multiplication, division, minimum, maximum, powers, etc., together would constitute an extension to naïve interval arithmetic that can begin to account for dependence between inputs.

The table above gives formulas for single arithmetic sums. For example, suppose the dependence relation between $A = [0,1]$ and $B = [1,11]$ is of the form $D(r = -0.5)$ as depicted in *Figure 1*, then the sum $A + B$ is surely within $[\text{env}(w(A, -r)+b_1, a_1+w(B,-r)), \text{env}(a_2+w(B,1+r),w(A,1+r)+b_2)]$ = $[\text{env}(w([0,1], 0.5)+1, 0+w([1,11], 0.5)), \text{env}(1+w([1,11],1-0.5),w([0,1],1-0.5)+11)]$ = $[\text{env}(0.5+1, 0+6), \text{env}(1+6, 0.5+11)]$ = $[[1.5,6], [7,11.5]]$ = $[1.5, 11.5]$. This interval is an improvement to both bounds over $[1,12]$ obtained by standard interval analysis that does not consider their dependence. The bounds accounting for this kind of dependence will be tighter than $[a_1+b_1, a_2+b_2]$ whenever r is less than zero. Another example involves a special case of the $D(r)$ dependence family which is the opposite dependence relation $O = D(-1)$. If A and B have this dependence, then their sum $A+B$ is sure to be within $[2,11]$. The tighter result arises in this case because the possible pairs of values from the two intervals are restricted to single combinations:

$a \in A$	$b \in B$	$a+b$
0	11	11,
⋮	⋮	⋮
0.1	10	10.1,
⋮	⋮	⋮
0.2	9	9.2,
⋮	⋮	⋮
0.3	8	8.3,
⋮	⋮	⋮
0.8	3	3.8,
⋮	⋮	⋮
0.9	2	2.9,
⋮	⋮	⋮
1	1	2.

For this reason, the formula for addition under opposite dependence simplifies to $\text{env}(a_1+b_2, a_2+b_1)$ as shown in the table.

There is an important caveat about the difficulty of the deriving formulas for arithmetic functions for different dependence relations. Monotonicity of the dependence relation does not ensure that the bounds on an arithmetic function can be found by testing two endpoints. Consider bounding the addition of intervals $A = [3,5]$ and $B = [4,11]$ that have a parabolic dependence defined by the constraint $(u - 1)^2 = v$. This corresponds to $\lambda = 1$, $q = 0$, and $r = -2$ atan($\frac{1}{2}$)/$\pi \approx -0.295$ and is the left branch of a parabola, so the dependence is monotone. (It is depicted as the decreasing curve in the third graph from the left in Figure 4.) The endpoints of the dependence relation might seem to suggest that the bounds on the sum would be $a+b = 3+11 = 14$ and $a+b = 5+4 = 9$. But the minimal value of the sum is actually obtained from the combination of $a = 4^5/7$ with $b = 4^1/7$, which is $8^6/7 \approx 8.857$. The values correspond to $u = {}^6/7$ and $v = {}^1/49$. This example shows that even when the dependence is a one-pair relation that is a monotone function, even the simplest arithmetic function, addition, cannot be evaluated by enveloping the results at the endpoints. Inspection of the endpoints or corners of the dependence relation only generally suffices to find the bounds on the arithmetic function if the edges of the dependence relation are straight lines and the arithmetic function is addition.

Accounting for dependence can sometimes lead to substantial numerical improvements over interval calculations that make no account of dependence. Although they are generally modest for addition, they can be large for other mathematical operations. For instance, if $A = [0,1]$ and $B = [1,11]$ have the opposite dependence relation O, the range of their product $A \times B$ is $[0,3.025]$, which is only a third of the width of the interval $[0,11]$ obtained by the standard calculation.

5. Uncertainty about the dependence

Specifying the dependence relations between input intervals is the prerogative and responsibility of the analyst. They should represent available information about constraints between the inputs. Because the specification of such dependencies is a matter of engineering judgment or empirical evidence, there may be uncertainty about how it should be done. In particular, for instance, one may not be able to ascribe a precise value to a correlation coefficient r. In such cases, it might be reasonable to use an interval to characterize r. The bounds on an arithmetic function of intervals in this case can be found by taking the union (or convex hull) of bounds obtained under each possible correlation coefficient within the interval.

When one does not know anything about the dependence at all, the all-pairs dependence relation we call nondependence should be used. This reduces all arithmetic calculations to the traditional interval formulas. This choice allows an analyst to compute *conservative* answers that enclose all possible results. Such a simple strategy is not available in probability theory. Assuming independence (or, indeed, any dependence) between random variables would not allow one to

find the bounds on an arithmetic function when their dependence is unknown. To do this, one must resort to computing the Fréchet convolutions (Ferson et al. 2004). This difference shows that nondependence is not really analogous to independence as it is recognized by probabilists. Although interval researchers often refer to nondependence as independence, and nondependence has sometimes been considered a kind of independence (Couso et al. 2000; cf. Ferson et al. 2004), we think that they are such distinct ideas that special care should be made to distinguish between them.

6. Multivariate dependence relations

So far, we have discussed only bivariate dependence relations, but there are multivariate generalizations as well. For example, $D_3 \subseteq [0,1] \times [0,1] \times [0,1]$ is a trivariate dependence relation if it contains at least one triple for every marginal value. Likewise, $D_k \subseteq [0,1]^k$ is a *k*-dimensional dependence relation if it contains at least one element for every marginal value. We can denote the set of all possible *k*-dimensional dependence relations as $\mathcal{D}_k$. We have been calling $\mathcal{D}_2$ simply $\mathcal{D}$. There is a *k*-dimensional generalization of P, but not of O.

The problem of accounting for dependencies among intervals in complex mathematical expressions may be much more difficult than it is for the binary operations considered in this note. Strategies for conveniently calculating best possible bounds await development. It may be difficult to properly handle such calculations as a sequence of binary operations on intervals. For example, suppose $A = [2,4]$, $B = [4,7]$, and $C = [3,9]$, where A and C have the opposite dependence relation O, and that the mathematical expression to be evaluated is $AB+C$. Approached as a composition of binary operations, the calculation would need to evaluate AB first and only then the sum. However, the information about the dependence between A and C is inaccessible once the multiplication occurs. What does dependence information about two variables imply about the dependence between functions of these variables? Simple simulations show that the best possible bounds on the function $AB+C$ given the opposite dependence between A and C are [17,31]. This interval can be obtained by assuming opposite dependence between C and the product AB, but it is not clear that assumptions like this are always permissible, or, in general, what theory governs dependence in interval calculations.

7. Conclusions

This paper has introduced the notion of dependence within interval calculations. Dependence is defined to be any restriction on the possible pairings of values from the respective intervals. Such

restrictions can be modeled as subsets of the unit square, which are relations (rather than functions) between the margins of a multivariate interval. As copulas abstract the notion of dependence out of joint distributions in probability theory, these structures extract the dependence out of multivariate intervals.

We have derived some exemplary formulas for bounding the results of interval addition under a handful of possible dependence relations, but the general computational problem of accounting for dependencies among intervals in arbitrary interval computations remains largely unstudied. Dependence information about two interval variables does not necessarily imply the dependence between functions of these variables. Further work is necessary to develop and implement convenient algorithms to enable routine calculations that take account of dependence among intervals. Further work is also needed to explore the role that conditionalization might play in the context of interval dependence.

References

Chernousko, F.L. 1988. *Estimation of the Phase Space of Dynamical Systems.* Nauka Publishers, Moscow, (in Russian).

Chernousko, F.L. 1994. *State Estimation for Dynamical Systems.* CRC Press, Boca Raton, Florida.

Couso, I., D. Moral and P. Walley. 2000. A survey of concepts of independence for imprecise probabilities. *Risk Decision and Policy* 5: 165-181.

Ferson, S., W. Troy Tucker and W.L. Oberkampf. 2004. The notion of independence when probabilities are imprecise. 9th ASCE EMD/SEI/GI/AD Joint Specialty Conference on Probabilistic Mechanics and Structural Reliability (PMC2004), Albuquerque, New Mexico.

Ferson, S., R.B. Nelsen, J. Hajgos, D.J. Berleant, J. Zhang, W.T. Tucker, L.R Ginzburg, and W.L. Oberkampf. 2004. *Dependence in Probabilistic Modeling, Dempster-Shafer Theory, and Probability Bounds Analysis*. SAND2004-3072, Sandia National Laboratories, Albuquerque, NM. http://www.ramas.com/depend.pdf

Kreinovich, V., J. Beck and H.T. Nguyen. 2005. Ellipsoids and ellopspod-shaped fuzzy sets as natural multi-variate generalizations of intervals and fuzzy numbers: how to elicit them from users, and how to use them in data processing. *Information Sciences* [to appear]. http://www.cs.utep.edu/vladik/2005/tr05-13.pdf

Kreinovich, V., J. Hajagos, L.R. Ginzburg and S. Ferson. 2006 [tentative]. Propagating uncertainty through quadratic approximations to complex models. SAND2006-xxxx, Sandia National Laboratories, Albuquerque, NM. http://www.ramas.com/quadratic.pdf.

Nelsen, R.B. 1999. *An Introduction to Copulas*. Lecture Notes in Statistics 139, Springer-Verlag, New York.