Handling dependencies between variables with imprecise probabilistic models.

Sebastien Destercke & Eric Chojnacki
Institut de Radioprotection et de Sûreté Nucléaire, BP3, 13115, St Paul-lez-Durance, Cadarache, France

ABSTRACT: Two problems often encountered in uncertainty processing (and especially in safety studies) are the following: modeling uncertainty when information is scarce or not fully reliable, and taking account of dependencies between variables when propagating uncertainties. To solve the first problem, one can model uncertainty by sets of probabilities rather than by single probabilities, resorting to imprecise probabilistic models. Iman and Conover method is an efficient and practical means to solve the second problem when uncertainty is modeled by single probabilities and when dependencies are monotonic. In this paper, we propose to combine these two solutions, by studying how Iman and Conover method can be used with imprecise probabilistic models.

1 INTRODUCTION

Modeling available information about input variables and propagating it through a model are two main steps of uncertainty studies. The former step consists in choosing a representation fitting our current knowledge or information about input variables or parameters, while the latter consists in propagating these information through a model (here functional) with the aim to estimate the uncertainty on the output(s) of this model. In this paper, we consider that uncertainty bears on N variables $X_1, \ldots, X_N$ defined on the real line. For all $i = 1, \ldots, N$, we note x_i a particular value taken by X_i.

Sampling methods such as Monte-Carlo sampling or Latin Hypercube sampling (Helton and Davis 2002) are very convenient tools to simulate and propagate random variables $X_1, \ldots, X_N$. Most of the time, they consists in sampling M realizations $(x_1^j, \ldots, x_N^j),\ j = 1, \ldots, M$ of the N random variables, thus building a $M \times N$ sample matrix S. Each line of the matrix S can then be propagated through a model $T : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$. When using such sampling technics, it is usual to assume:

1. That uncertainty on each X_i is representable by a unique probability density p_i associated to a unique cumulative distribution F_i, with
$$F_i(x) = P_i([-\infty, x]) = \int_{-\infty}^{x} p_i(x)dx.$$
2. That variables $X_1, \ldots, X_N$ are independent, that is that their joint probability distribution is provided by the product of the marginal probability distributions.

In real applications, both assumptions can be challenged in a number of practical cases: the first when available information is scarce, imprecise or not fully reliable, and the second when independence between variables cannot be proved or is clearly unrealistic. As shown in (Ferson and Ginzburg 1996), making such assumptions when they are not justified can lead to underestimations of the final uncertainty on the output, possibly leading to bad decisions.

Although there exist some very practical solutions to overcome either scarceness of the information or dependencies between variables, there are not a lot of methods treating both problems at the

same time. In this paper, we propose and discuss such a method, that combines the use of simple imprecise probabilistic representations with classical technics used to model monotonic dependencies between variables (namely, Iman and Conover method). The paper is divided in two main sections: section 2 is devoted to basics needed to understand the paper, and section 3 explains and discusses the proposed method.

2 PRELIMINARIES

This section recalls the main principles of Iman and Conover (Iman and Conover 1982) method to integrate monotonic dependencies in a sampling matrix and introduces possibility distributions (Baudrit and Dubois 2006) and probability boxes (p-boxes for short) (Ferson, Ginzburg, Kreinovich, Myers, and Sentz 2003), the two practical probabilistic models we are going to consider. More details can be found in the references.

2.1 Integrating monotonic dependencies in sampling procedures

The first problem we deal with is the integration of dependencies into sampling schemes. In the sequel, $S_{i,j}$ denote the matrix element in the i^{th} line and j^{th} column of S, while $S_{.,j}$ and $S_{i,.}$ respectively denote the j^{th} column and i^{th} line of S.

Suppose we consider two variables X, Y and a sample (x_j, y_j) of size M of these two variables. Then, if we replace the values x_j and y_j by their respective ranks (The lowest value among x_j receive rank 1, second lowest rank 2, ..., and similarly for y_j), their spearman rank correlation coefficient r_s, which is equivalent to the Pearson correlation computed with ranks, is given by

$$r_s = 1 - \left(\frac{6 \sum_{j=1}^{M} d_j^2}{M(M^2 - 1)} \right)$$

with d_j the difference of rank between x_j and y_j. Spearman correlation r_s have various advantages:

i it allows to measure or characterize monotonic (no necessarily linear) dependencies between variables

ii it depends only on the ranks, not on the particular values of the variables (i.e. it is distribution-free).

Although Spearman correlations rank are not able to capture all kinds of dependencies, they remain nowadays one of the best way to elicit dependency structures (Clemen, Fischer, and Winkler 2000).

Given a sample matrix S and a $N \times N$ target rank correlation matrix R (e.g. elicited from experts), Iman and Conover (Iman and Conover 1982) propose a method to transform the matrix S into a matrix S^* such that the rank correlation matrix R^* of S^* is close to the target matrix R. This transformation consists in re-ordering the elements in each column $S_{.,j}$ of S, without changing their values in S, so that the result is the matrix S^*. The transformation consists in the following steps:

1. Build a $M \times N$ matrix W whose N columns are random re-orderings of the vector $(a_1, \ldots, a_M)$, where $a_i = \phi^{-1}(i/(M+1))$, ϕ^{-1} being the inverse of a standard normal cumulative distribution, that is

$$\forall x,\ \phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2} du \right).$$

   Let C be the $N \times N$ correlation matrix associated to W.

2. Build a lower triangular $N \times N$ matrix G such that $G\ C\ G' = R$ with G' the transpose of G. This can be done in the following way: use Cholesky factorization procedure to decompose C and R into $C = C_\triangle\ C'_\triangle$ and $R = R_\triangle\ R'_\triangle$, with both $C_\triangle, R_\triangle$ lower triangular matrix (due to the fact that correlation matrices C, R are, by definition, positive definite and symmetric). Then, G is given by $G = R_\triangle\ C_\triangle^{-}1$ and the transpose follows. Note that G is still a lower triangular matrix.

3. Compute the $M \times N$ matrix $W^* = W\ G'$.

4. In each column $S_{.,j}$ of the original sample matrix, re-order the sampled values so that they are ranked as in the column $W^*_{.,j}$, thus obtaining a matrix S^* whose rank correlation matrix R^* is close to R (but not forcefully equal, as for a given number M of samples, rank correlations coefficients can only assume a finite number of distinct values).

This method allows to take account of monotonic dependencies between the variables in sampling schemes (and, therefore, in the subsequent propagation), without making any assumptions about the shape of probability distributions and

without changing the sampled value (it just rearranges their pairings in the sample matrix). It is also mathematically simple and applying it do not require complex tools, as would other approaches involving, for example, copulas (Nelsen 2005).

2.2 *Modeling uncertainty with sets of probabilities*

The second problem concerns situations where available information is scarce, imprecise or not fully reliable. Such information can come, for instance, from experts, from few experimental data, from sensors, etc. There are many arguments converging to the fact that, in such situations, a single probability distribution is unable to account for the scarcity or imprecision present in the available information, and that such information would be better modeled by sets of probabilities (see (Walley 1991, Ch.1) for a summary and review of such arguments).

Here, we consider two such models: p-boxes and possibility distributions. They are both popular, simple and are instrumental to represent or elicit information from experts (for more general models and longer discussion, see (Destercke, Dubois, and Chojnacki 2007)).

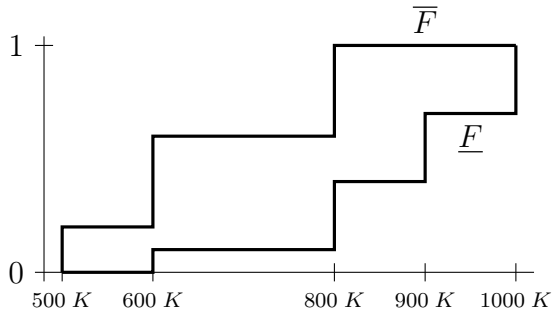

Figure 1: Illustration of a p-box

P-boxes (short name for probability boxes) are the imprecise counterparts of cumulative distributions. They are defined by an upper ($\overline{F}$) and a lower ($\underline{F}$) cumulative distributions forming a pair $[\underline{F}, \overline{F}]$ describing the uncertainty: the information only allows us to state that the true cumulative distribution is between $\underline{F}$ and $\overline{F}$, and any cumulative distribution F such that $\underline{F} \leq F \leq \overline{F}$ is coherent with the available information. A p-box induces a set $\mathcal{P}_{[\underline{F},\overline{F}]}$ of probabilities, such that

$$\mathcal{P}_{[\underline{F},\overline{F}]} = \{P | \forall x \in \mathbb{R},\ \underline{F}(x) \leq P([-\infty, x]) \leq \overline{F}\}.$$

P-boxes are appropriate models when experts provide a set of (imprecise) percentiles, when considering the error associated to sensor data, when we have only few experimental data or when we have only information about some characteristics of a distribution (Ferson, Ginzburg, Kreinovich, Myers, and Sentz 2003). Consider the following expert opinion about the temperature of a fuel rode in a nuclear reactor core during an accidental scenario:

- Temperature is between 500 and 1000 K
- The probability to be below 600 K is between 10 and 20%
- The probability to be below 800 K is between 40 and 60%
- The probability to be below 900 K is between 70 and 100%

Figure 1 illustrates the p-box resulting from this expert opinion.

Possibility distributions correspond to information given in terms of confidence intervals, and thus correspond to a very intuitive notion. A possibility distribution is a mapping $\pi : \mathbb{R} \to [0, 1]$ such that there is at least one value x for which $\pi(x) = 1$. Given a possibility distribution π, possibility Π and necessity N measures of an event A are respectively defined as:

$$\Pi(A) = \max_{x \in A} \pi(x) \text{ and } N(A) = 1 - \pi(A^c)$$

with A^c the complement of A. For any event A, $N(A) \leq \Pi(A)$, and possibility and necessity measure are respectively interpreted as upper and lower confidence levels given to an event, defining a set of probabilities $\mathcal{P}_\pi$ (Dubois and Prade 1992) such that

$$\mathcal{P}_\pi = \{P | \forall A \subseteq \mathbb{R} N(A) \leq P(A) \leq \Pi(A)\}$$

with P a probability distribution. For a given possibility distribution π and for a given value $\alpha \in [0, 1]$, the (strict) α-cut of π is defined as the set

$$\pi_\alpha = \{x \in \mathbb{R} | \pi(x) > \alpha\}.$$

Note that α-cuts are nested (i.e. for two values $\alpha < \beta$, we have $\pi_\beta \subset \pi_\alpha$). An α-cut can then be interpreted as an interval to which we give confidence $1 - \alpha$ (The higher α, the lower the confidence). α-cuts and the set of probabilities $\mathcal{P}_\pi$ are related in the following way

$$\mathcal{P}_\pi = \{P | \forall \alpha \in [0, 1],\ P(\pi_\alpha) \geq 1 - \alpha\}.$$

Possibility distributions are appropriate when experts express their opinion in term of nested confidence intervals or more generally when information is modeled by nested confidence intervals (Baudrit, Guyonnet, and Dubois 2006). As an example, consider an expert opinion, still about the temperature of a fuel rode in a nuclear reactor core, but this time expressed by nested confidence intervals:

- Probability to be between 750 and 850 K is at least 10%
- Probability to be between 650 and 900 K is at least 50%
- Probability to be between 600 and 950 K is at least 90%
- Temperature is between 500 and 1000 K (100% confidence)

Figure 2 illustrates the possibility distribution resulting from this opinion.

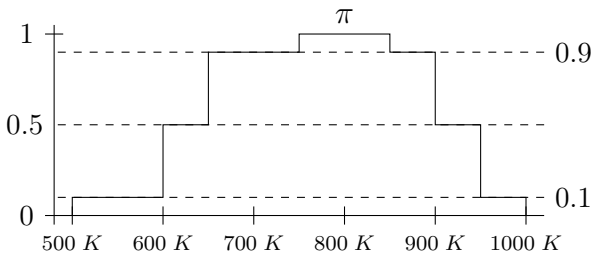

Figure 2: Illustration of a possibility distribution

3 PROPAGATING WITH DEPENDENCIES AND IMPRECISE MODELS

Methods presented in Section 2 constitute very practical solutions to solve two different problems often encountered in applications. As both problems can be encountered in a same application, it would be interesting to blend these two tools. Such a blending is proposed in this section.

3.1 *Sampling with imprecise probabilistic models*

When uncertainty on a (random) variable X is modeled by a precise cumulative distribution F_X, then simulating this variable X by sampling methods usually consists of drawing values α coming from a uniform law on $[0,1]$, and then to associate the (precise) value $F^{-1}(\alpha)$ to each value α (see Figure 3.A). In the case of a N-dimensional problem simulated by M samples, the j^th sample consists of a vector $(\alpha_1^j, \ldots, \alpha_N^j)$, to which is associated the realization $(F^{-1}(\alpha^j)_1, \ldots, F^{-1}(\alpha^j)_N) = (x_1^j, \ldots, x_N^j)$. Let us now detail what would be the result of such a sampling with imprecise models.

P-boxes: since a p-box is described by (lower and upper) bounds on cumulative distributions, to each value α do not longer correspond a unique inverse value, but a set of possible values. This set of possible values correspond to the interval bounded by the upper ($\overline{F}^{-1}$) and lower ($\underline{F}^{-1}$) pseudo inverses, defined, for all $\alpha \in (0,1]$ as follows:

$$\overline{F}^{-1} = \sup\{x \in \mathbb{R} | \overline{F}(x) < \alpha\}$$

$$\underline{F}^{-1} = \inf\{x \in \mathbb{R} | \underline{F}(x) > \alpha\}$$

See Figure 3.B for an illustration. Thus, given a p-box $[\underline{F}, \overline{F}]$, to a sampled value $\alpha \in [0,1]$ we associate the interval Γ_α such that

$$\Gamma_\alpha := [\overline{F}^{-1}(\alpha), \underline{F}^{-1}(\alpha)]$$

Possibility distributions: In the case of a possibility distributions, it is natural to associate to each value α the corresponding α-cut (see Figure 3.C for illustration). Anew, this α-cut π_α is, in general, not a single value but an interval.

We can see that, by admitting imprecision in our uncertainty representation, usual sampling methods do not longer provide precise values but intervals (which are effectively the imprecise counterpart of single values). With such models, elements of matrix S can be intervals and propagating them through a model T will require to use interval analysis technics (Moore 1979). Although achieving such a propagation is more difficult than single point propagation when the model T is complex, it can still remain tractable, even for high dimensional problems (see (Oberguggenberger, King, and Schmelzer 2007) for example). Nevertheless, propagation is not our main concern here, and sampling scheme can be considered independently of the subsequent problem of propagation.

Also note that above sampling procedures have been considered by Alvarez (Alvarez 2006) in the more general framework of random sets, of which p-boxes and possibility distributions constitute two particular instances. Let us now see how Iman and Conover method can be extended to such models.

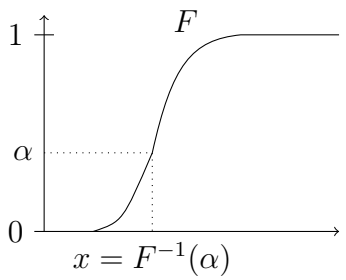

Fig. 3.A: precise prob.

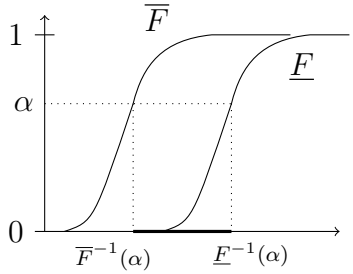

Fig. 3.B: p-box

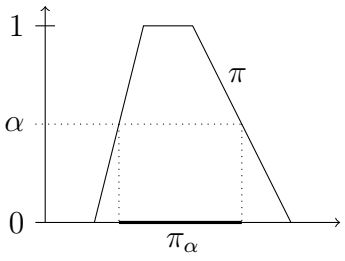

Fig. 3.C: possibility dist.

Figure 3: Sampling from precise and imprecise probabilistic models: illustration

3.2 *Extension of Iman and Conover method*

We first recall some notions coming from order theory. Let P be a set and $\leq$ a relation on the elements of this set. Then, $\leq$ is a **complete partial order** if it is reflexive, antisymmetric and transitive, that is if for all triplet a, b, c of elements in P

$$a \leq a \text{ (reflexivity)} \tag{1}$$

$$\text{if } a \leq b \text{ and } b \leq a, \text{ then } a = b \text{ (antisymmetry)} \tag{2}$$

$$\text{if } a \leq b \text{ and } b \leq c, \text{ then } a \leq c \text{ (transitivity)} \tag{3}$$

and if for two elements a, b, neither $a \leq b$ nor $b \leq a$, then a and b are said to be incomparable. A partial order $\leq$ is total, and is called an order (or a linear order), if for every pair a, b in P, we have either $a \leq b$ or $b \leq a$.

When uncertainty is modeled by precise probabilities, sampled values are precise, and the main reason for being able to apply Iman and Conover method in this case is that there is a natural complete ordering between real numbers, and that to any set of values corresponds a unique ranking. This is no longer the case when realizations are intervals, since in most cases only partial orderings can be defined on sets of intervals (due to the fact that they can be overlapping, nested, disjoint,...). Given two intervals $[a, b], [c, d]$, it is common to consider the partial ordering such that $[a, b] < [c, d]$ if and only if $b < c$, and to consider that two intervals are incomparable as soon as they overlap. This partial order is commonly called interval order. Adapting Iman and Conover method when samples are general intervals thus seems difficult and would result in a not very convenient tool, since one would have to consider every possible extension of the partial ordering induced by the interval ordering.

To circumvent this problem and to be able to apply Iman and Conover method in an easy way on p-boxes and possibility distributions, we have to define a complete ordering on the elements sampled from these two representation.

First, note that when uncertainty on a variable X is modeled by a single (invertible) cumulative distribution F_X, there is a one-to-one correspondence between the ranking of sampled values $\alpha^j \in [0, 1]$ and the ranking of corresponding values of X, in the sense that, for two values α^i, α^j, we have

$$\alpha^i < \alpha^j \iff F^{-1}(\alpha^i) < F^{-1}(\alpha^j). \tag{4}$$

We will use this property to extend Iman and Conover technics when realizations are either intervals Γ_α coming from a p-box or α-cuts π_α of a possibility distribution.

P-boxes: Consider first a p-box $[\underline{F}, \overline{F}]$. For such a p-box, the ordering similar to Equation (4) means that for two values α, β in $[0, 1]$, we have $\alpha < \beta \rightarrow \Gamma_\alpha < \Gamma_\beta$, which is equivalent to impose a complete ordering between "cuts" Γ of the p-box. We note this ordering $\leq_{[\underline{F},\overline{F}]}$. Given two intervals $[a, b], [c, d]$, this definition is equivalent to state that an interval $[a, b] \leq_{[\underline{F},\overline{F}]} [c, d]$ if and only if $a \leq c$ and $b \leq d$. Roughly speaking, taking such an ordering means that the rank of an interval increases as it "shifts" towards higher values. Note that the ordering $\leq_{[\underline{F},\overline{F}]}$ is complete only when intervals are sampled from a p-box, and that incomparability can appear in more general cases (e.g. when intervals are nested, or when they come from general random sets).

Possibility distributions: given a possibility distribution π, the ordering similar to Equation (4) is equivalent to consider a complete ordering on α-cuts induced by inclusion: for two values α, β in $[0, 1]$, we have $\alpha < \beta \rightarrow \pi_\alpha \supset \pi_\beta$. We note $\leq_\pi$ the ordering such that

$$[a, b] \leq_\pi [c, d] \text{ if and only if } [a, b] \supset [c, d]$$

with $[a, b]$, $[c, d]$ two intervals. Here, the rank of an interval increases as it gets more precise (narrower). Again, the ordering $\leq_\pi$ is complete only when intervals are sampled from a possibility distribution.

Now that we have defined complete orderings on intervals sampled from p-boxes and possibility distributions, we can apply Iman and Conover method without difficulty to these models. Nevertheless, one must pay attention that a same value of rank correlation will have different meaning and interpretation, depending on the chosen representation (and, consequently, on the chosen ordering).

In the case of p-boxes defined on multiple variables, a positive (negative) rank correlation always means that to higher values are associated higher (lower) values. The main difference with single cumulative distributions is that samples are now intervals instead of single values. In the case of p-boxes, the application of Iman and Conover method can then be seen as a "simple" extension of the usual method, with the benefits that imprecision and scarceness of information is now acknowledged in the uncertainty model. As a practical tool, it can also be seen as a means to achieve a robustness study (concerning either distribution shape or correlation coefficients). Since correlation coefficients can seldom be exactly known and are often provided by experts, such a robustness interpretation appears appealing. Also note that the ordering $\leq_{[\underline{F},\overline{F}]}$ is a refinement of the classical ordering considered on intervals, and reduce to the classical ordering between numbers when samples are single values. All this indicates that using Iman and Conover method on p-boxes is also equivalent to induce monotonic dependencies between variables.

Contrary to p-boxes, rank correlations related to possibility distributions and to the ordering $\leq_\pi$ cannot be considered as an extension of the classical Spearman rank correlations. To see this, simply note that the ordering $\leq_\pi$, based on inclusion between sets, is not a refinement of the classical ordering considered on intervals, and do not reduce to the classical ordering of numbers when samples are single values. In the case of possibility distributions, a positive (negative) rank correlation between two variables X, Y means that to more precise descriptions of the uncertainty on X will be associated more (less) precise descriptions of the uncertainty on Y, i.e. that to narrower intervals will correspond narrower (broader) intervals. Such dependencies can be used when sensors or experts are likely to be correlated, or in physical models where knowing a value with more precision means knowing another one with less precision (of which Heisenberg principle constitutes a famous example). Such kind of dependencies has poor relation with monotonic dependencies, meaning that using the proposed extension to possibility distribution is NOT equivalent to assume monotonic dependencies between variables, but rather to assume a dependency between the precision of the knowledge we have on variables. Nevertheless, if monotonic dependencies have to be integrated and if information is modeled by possibility distributions, it is always possible to extract a corresponding p-box from a possibility distribution, and then to sample from this corresponding p-box (see (Baudrit and Dubois 2006)).

4 CONCLUSIONS

Integrating known correlation between variables and dealing with scarce or imprecise information are two problems that coexist in many real applications. The use of rank correlation through the means of Iman and Conover method and the use of simple imprecise probabilistic models are practical tools to solve these two problems. In this paper, we have proposed an approach to blend these two solutions, thus providing a practical tool to cope (at the same time) with monotonic dependencies between variables and with scarceness or imprecision in the information.

Sampling methods and complete orderings related to possibility distributions and p-boxes have been studied and discussed. They allow to apply Iman and Conover method to these two models without additional computational difficulties. We have argued that, in the case of p-boxes, rank correlations can still be interpreted in terms of monotonic dependencies, thus providing a direct extension of Iman and Conover method, with the advantage that it can be interpreted as an integrated robustness study. The interpretation concerning

possibility distributions is different, as it is based on set inclusion, and describes some dependencies between the precision of the knowledge we can acquire on different variables. We suggest that such correlation can be useful in some physical models, or when sources of information (sensors, experts) are likely to be correlated.

In our opinion, the prime interest of the suggested extensions is practical, as they allow to use very popular and efficient numerical technics such as Latin Hyper Cube Sampling and Iman and Conover method with imprecise probabilistic models. Moreover, the proposed extensions can benefits from all the results concerning these numerical technics (for instance, see (Sallaberry, Helton, and Hora 2006)).

References

Alvarez, D. A. (2006). On the calculation of the bounds of probability of events using infinite random sets. *I. J. of Approximate Reasoning 43*, 241–267.

Baudrit, C. and D. Dubois (2006). Practical representations of incomplete probabilistic knowledge. *Computational Statistics and Data Analysis 51*(1), 86–108.

Baudrit, C., D. Guyonnet, and D. Dubois (2006). Joint propagation and exploitation of probabilistic and possibilistic information in risk assessment. *IEEE Trans. Fuzzy Systems 14*, 593–608.

Clemen, R., G. Fischer, and R. Winkler (2000, August). Assessing dependence : some experimental results. *Management Science 46*(8), 1100–1115.

Destercke, S., D. Dubois, and E. Chojnacki (2007). Relating practical representations of imprecise probabilities. In *Proc. 5th Int. Symp. on Imprecise Probabilities: Theories and Applications*.

Dubois, D. and H. Prade (1992). On the relevance of non-standard theories of uncertainty in modeling amd pooling expert opinions. *Reliability Engineering and System Safety 36*, 95–107.

Ferson, S., L. Ginzburg, V. Kreinovich, D. Myers, and K. Sentz (2003). Constructing probability boxes and dempster-shafer structures. Technical report, Sandia National Laboratories.

Ferson, S. and L. R. Ginzburg (1996). Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety 54*, 133–144.

Helton, J. and F. Davis (2002). Illustration of sampling-based methods for uncertainty and sensitivity analysis. *Risk Analysis 22*(3), 591–622.

Iman, R. and W. Conover (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics 11*(3), 311–334.

Moore, R. (1979). *Methods and applications of Interval Analysis*. SIAM Studies in Applied Mathematics. Philadelphia: SIAM.

Nelsen, R. (2005). Copulas and quasi-copulas: An introduction to their properties and applications. In E. Klement and R. Mesiar (Eds.), *Logical, Algebraic, Analytic, and Probabilistics Aspects of Triangular Norms*, Chapter 14. Elsevier.

Oberguggenberger, M., J. King, and B. Schmelzer (2007). Imprecise probability methods for sensitivity analysis in engineering. In *proc. of the 5th Int. Symp. on Imprecise Probabilities: Theories and Applications*, pp. 317–326.

Sallaberry, C., J. Helton, and S. Hora (2006). Extension of latin hypercube samples with correlated variables. Tech. rep. sand2006-6135, Sandia National Laboratories, Albuquerque. http://www.prod.sandia.gov/cgi-bin/techlib/access-control.pl/2006/066135.pdf.

Walley, P. (1991). *Statistical reasoning with imprecise Probabilities*. New York: Chapman and Hall.