Uncertainties Quantification and Propagation of Multiple Correlated Variables with Limited Samples

Zhanpeng Shen¹,², Xueqian Chen¹, Xinen Liu¹, Chaoping Zang²

¹ Institute of Systems Engineering, China Academy of Engineering Physics, Mianyang 621999, China.
² College of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract. In order to estimate the reliability of an engineering structure based on limited test data, it is distinctly important to address both the epistemic uncertainty from lacking in samples and correlations between input uncertain variables. Both the probability boxes theory and copula function theory are utilized in proposed method to represent uncertainty and correlation of input variables respectively. Moreover, the uncertainty of response of interest is obtained by uncertainty propagation of correlated input variables. Nested sampling technique is adopted here to insure the propagation is always feasible and the response’s uncertainty is characterized by a probability box. Finally, a numerical example illustrates the validity and effectiveness of our method. The results indicate that the epistemic uncertainty cannot be conveniently ignored when available samples are very limited and correlations among input variables may significantly affect the uncertainty of responses.

1. Introduction

Modeling and simulation have been heavily used in engineering systems, especially under the condition of quite limited testing available. Due to great expense, undeveloped technique or safety problem, et al., full scale testing may not be feasible or only very limited number of tests are available under actual use environments for some complex systems, such as crashing car, aircraft in the atmosphere, nuclear power reactors, and the nuclear weapon stockpile [1]. The reliability assessment of these systems relies on simulation more strongly than experiments. As a result, high fidelity model is suggested and uncertainties in the model are desired to be quantified. Generally, the latter, i.e. uncertainty quantification (UQ), is much more important and practical in engineering especially when the fidelity of model has been on a high level.

Input variables of model usually are determined by limited testing of material, components, or subsystems. Due to the cost of tests, the samples from testing results are too sparse to estimate the statistics parameters of variables accurately. It means the uncertainty of variables could not be quantified exactly with such limited test data [2], and so-called “epistemic uncertainty” has to be taken into account further. In contrast with aleatory uncertainty which arises from an inherent randomness of the variables and cannot be reduced, epistemic uncertainty derives from a lack of knowledge about the behavior of the system and it would decrease to zero as more and more tests or studies are performed [3]. A variety of alternative approaches have been developed to represent uncertainty of variables, such as mathematical statistics, interval analysis [4], D-S evidence theory [5, 6], fuzzy theory [7],
possibility theory [8], Bayesian inferences [9, 10], et al. Some of them were compared with each other in Ref. [11]. A recommended approach, probability boxes (i.e. p-boxes) method [12-14], could characterize both aleatory and epistemic uncertainty at the same time and it will be introduced in Section 2.1 in detail.

In order to estimate the reliability of an engineering structure or system, the process named uncertainty propagation should be performed to calculate the uncertainty of system response quantities (SRQs) of interest according to uncertainties of input variables. In the case of simple mathematical models, Taylor series expansions can be used to approximate the model and the perturbation method can be adopted to obtain uncertainty of SRQs of interest. However, for complex models, it would be difficult to use the Taylor series approximation, and more advanced techniques are needed. Sampling is a popular and tremendously useful method to propagate uncertainties for complex model [11, 15, 16]. Since the model is treated as a black box regardless of its complexity, uncertainty propagation could be always carried out by sampling method. Probability bounds analysis (PBA) [12], which is also recommended, can obtain the responses’ uncertainties represented with p-box using a nested sampling technique based upon input variables’ uncertainties.

There are two cases in uncertainty propagation when the number of input variables is larger or equal than two. In first case that the input variables are independent with each other, uncertainty propagation would not vary in essence from single variable to multiple variables. While in the other case which is extremely common in engineering, the input variables are dependent. Then the correlation between variables should be taken into account in uncertainty propagation, and it matters to the uncertainty of responses especially when uncertain variables are strongly correlated. Several approaches have been proposed to deal with the correlated variables [17-21]. In probability framework, copula function is commonly utilized to construct the joint distribution of correlated variables as the representation of aleatory uncertainty. The key problem, which is also the particularly difficult issue, is how to construct the accurate copula function to represent variables’ uncertainty when limited data is available. In other words, it is desired to represent both the epistemic uncertainty and correlations between the variables at the same time based on limited data.

This paper proposes a copula-based method for the process of uncertainty quantification and propagation with consideration of both the epistemic uncertainty from lacking in data and the correlation among variables. The p-box theory and copula functions are utilized in this paper and they are briefly introduced in Section 2. Then the detail process of our method is presented in Section 3. A numerical example is provided in Section 4 to verify the effectiveness of the proposed method and Section 5 summarizes the entire paper.

2. Technical background

2.1. Probability boxes

As mentioned above, uncertainty can be divided to two types: i) aleatory uncertainty that emphasize the natural variability or randomness in system and cannot be reduced by further empirical study (although it may be better characterized); ii) epistemic uncertainty that comes from scientific ignorance, measurement uncertainty, unobservability, censoring, or other lack of knowledge and can generally be reduced by additional empirical effort at least in principle.

Based on the idea that one can work with bounds on probability for the purpose of representing epistemic uncertainty within the context of probability theory, Williamson and Downs introduced interval-type bounds on cumulative distribution functions [13], which we call “probability boxes”, or “p-boxes” for short. Cumulative distribution function (CDF) is often used in probability theory to represent the probability $P(\xi \geq x)$, while the interval-value of CDF is provided in p-boxes theory with the nondecreasing functions $F_i(x)$ and $F_u(x)$. An illustration of a p-box is shown in Figure 1. If one fixes the variable $X$ to the value $x_i$, then an interval of probability of $\xi \geq x_i$ could be obtained as

$$F_i(x_i) \leq P(\xi \geq x_i) \leq F_u(x_i)$$

(1)
and the interval of probability of $\xi \leq x$ can be easily derived by the relation of $P(\xi \leq x) = 1 - P(\xi > x)$. Likewise, if one fixes the probability to $p_r$, then an interval of variable $\xi$ could also be obtained as

$$F_{u}^{-1}(p_r) \leq \xi \leq F_{l}^{-1}(p_r)$$

(2)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{An illustration of the p-box representing uncertain variable $X$}
\end{figure}

There are basically five ways to construct p-boxes: i) Direct assumption, ii) Modeling, iii) Resort to robust Bayes methods, iv) Constraint propagation, and v) Observation of measurements. The constructing approaches of p-box were deeply discussed in Ref. [14].

P-boxes provide convenient and comprehensive ways to handle several of the most practical serious problems faced by analysts, including imprecisely specified distributions, poorly known or even unknown dependencies, non-negligible measurement uncertainty, non-detects or other censoring in measurements, small sample size, inconsistency in the quality of input data, model uncertainty, and non-stationarity (non-constant distributions). In addition, Dempster-Shafer evidence theory has some similar advantages as p-boxes. Generally, a p-box can always be discretized to Dempster-Shafer evidence structure, but translating a D-S structure to a p-box is not an information-preserving operation. The connection of these two has been expounded in Ref. [14] in detail.

Sometimes we only know the upper and lower bounds of a p-box. However, sometimes we are acquainted with a large amount of distribution curves which constitute the p-box as shown in Figure 2. These distribution curves may be formed from distributions with a deterministic type but nondeterministic parameters, e.g. normal distribution functions with interval-valued means or interval-valued variances. Meanwhile they may also be mapped by the empirical cumulative distribution functions (ECDF) of the data from observation or calculation as shown in Figure 2. In any case, it should be emphasized that the curves include more information and provide more restriction to p-box than two bounds of p-box. Consequently, it is likely to overestimate the uncertainty of responses in propagation when only two bounds of p-box are used to quantify input variables uncertainty. Hence, nested sampling technique is adopted and all the distribution curves have been used to propagate uncertainty in this paper.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{A p-box consisting of 30 empirical CDFs each plotted with 50 points.}
\end{figure}
2.2. Copula function

The copula function is currently the most effective math tool in correlation analyses in the field of probability methods and is capable of connecting the joint distribution \( F(x_1, x_2, \ldots, x_d) \) of multiple random variables \( x_1, x_2, \ldots, x_d \) and their respective marginal distributions: \( F_1(x_1), F_2(x_2), \ldots, F_d(x_d) \) \[18\]. In the connection process, the copula function can consider the correlations among various random variables. Copula function was first proposed by Sklar in 1959 [22, 23] and was strictly defined in mathematics and developed by Nelsen [24, 25].

It is proved that given a joint distribution function \( F(x_1, x_2, \ldots, x_d) \) for marginal distributions \( F_1(x_1), F_2(x_2), \ldots, F_d(x_d) \), a copula function \( C \) exists for \( d \)-dimensional random variable \( x_1, x_2, \ldots, x_d \) to satisfy

\[
F(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))
\] (3)

It is suggested that copula is the function \( C(u_1, u_2, \ldots, u_d) \) defined from the domain \([0,1]^d\) to the codomain \([0,1]\). Some elementary properties of copulas are as follows:

a) \( C(u_1, u_2, \ldots, u_d) \) is monotonous and nondecreasing as each variable increases;
b) \( C(u_1, u_2, \ldots, 0, \ldots, u_d)=0 \) and \( C(1, \ldots, 1, u_1, \ldots, 1) = u_i; \)
c) If \( u_i \) (i=1, 2, ..., \( d \)) follows an standard uniform distribution, i.e. U(0,1), and they are independent with each other, then \( C(u_1, u_2, \ldots) = \prod_{i=1}^{d} u_i \), and the density function \( c(u_1, u_2, \ldots, u_d) \) would be equal to one.

The commonly used copula functions include the Gaussian copula function, the t-copula function and the Archimedean copula function. Based on the differences in its generator, the Archimedean copula function can be divided into the Gumbel copula function, the Clayton copula function and the Frank copula function. In this paper, Gaussian copula function was adopted to construct the joint distribution of multi-variables in the numerical example. The reader is referred to Refs. [18, 26] for more details of Archimedean copula.

The distribution function and density function of Gaussian copula with \( d \)-dimensional variables can be expressed as

\[
C_{G}(u_1, u_2, \ldots) = \Phi_{\rho}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_d))
\] (4)

\[
c_G(u_1, u_2, \ldots) = \frac{\partial^d C(u_1, u_2, \ldots)}{\partial u_1 \partial u_2 \ldots} = |\mathbf{p}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \zeta''(\mathbf{p}^{-1} - 1) \zeta \right)
\] (5)

where \( \mathbf{p} \) is the correlation coefficients matrix with 1-value diagonal elements, \( |\mathbf{p}| \) represents the determinant of \( \mathbf{p} \), \( \Phi_\rho \) is the multivariate standard normal distribution of \( d \)-dimensional variables with the correlation matrix \( \rho \), \( \Phi^{-1} \) represents the inverse function of standard normal distribution, \( \zeta''(\mathbf{p}^{-1} - 1) \) and \( \mathbf{I} \) is units matrix.

Similarly, t-copula distribution function in \( d \)-dimensions can be expressed as

\[
C_{T}(u_1, u_2, \ldots) = t_{\rho, k}(t^{-1}_k(u_1), t^{-1}_k(u_2), \ldots, t^{-1}_k(u_d))
\] (6)

where \( t_{\rho, k} \) is the multivariate t-distribution of \( d \)-dimensional variables with the correlation matrix \( \rho \) and \( k \) degrees of freedom. \( t^{-1}_k \) represents the inverse function of t-distribution with \( k \) degrees of freedom. The reader is referred to Refs. [28-31] to learn more details about these copulas.

The key to calculating the joint distribution function based copulas is the derivation of the parameters including marginal distribution parameters and correlation parameter among the variables. Common solution methods for parameters include the exact maximum likelihood (EML) method [32], the inference functions for margins (IFM) method [33] and the canonical maximum likelihood (CML) method [34]. All of them are based on maximum likelihood estimation. The EML method estimates all the parameters simultaneously. The IFM method estimates the marginal distribution parameters at first and deals with correlation parameters secondly. The CML method utilizes empirical distribution as the approximation of marginal distribution, and hence only correlation parameter needs to be estimated. Compared with the first two methods, the CML method has the advantage of permitting no awareness
of the variables’ marginal distributions types, but would be not accurate when sample size is small. Due to the availability at the situation of sparse samples and the compatibility with p-box theory, the IFM method is adopted in this paper.

3. Proposed Method

In this Section, both the uncertainties and correlation between input variables are quantified based on the sparse samples. Then uncertainty propagation is performed by PBA with the joint CDF of input variables constructed via copulas.

3.1. Quantification of uncertainties and correlation between the input variables

As mentioned above, the epistemic uncertainty due to lack of data samples is considered in this paper. The p-box theory is utilized to characterize both the aleatory and epistemic uncertainties of input variables. Assuming the probability distribution types of the variables are known, only the parameters need to be estimated and it becomes interval estimation instead of point estimation for representing the epistemic uncertainty. For example, the variable \( x \) is assumed to follow a normal distribution with unknown two parameters, i.e. the mean and variance, according to experience or historical data. And the two parameters should be estimated based on available samples. However, the epistemic uncertainty of the parameters due to lack of sufficient samples may be significant and should be taken in account, especially when the sample size is small. Thus, the parameters (i.e. mean and variance) of the normal distribution of \( x \) are interval-estimated (denoted by \([\mu_{\text{low}}, \mu_{\text{up}}]\) and \([\sigma^2_{\text{low}}, \sigma^2_{\text{up}}]\) respectively) instead of being point-estimated. As a result, a p-box representing variable \( x \)’s uncertainty is constructed by hundreds of normal CDF curves with means sampled form the interval \([\mu_{\text{low}}, \mu_{\text{up}}]\) and variance sampled from the interval \([\sigma^2_{\text{low}}, \sigma^2_{\text{up}}]\). It should be noted that the p-box is the combination of the marginal probability distributions of each variable in term of probability theory.

Corresponding to each curve in the p-box, a point estimation of the correlation parameter is acquired by IFM method. As the p-box consists hundreds of marginal probability distributions, there are hundreds of estimated values of the correlation parameter too. In fact, the dispersing of values is an existing evidence of epistemic uncertainty. As more data available are collected, the epistemic uncertainty would reduce, and the estimated range of the correlation parameter would narrow down.

Because the method proposed here does not rely on the number of input variables in essence leaving out the amount of calculation, the steps of the method are listed below in the situation of only two variables \( x \) and \( y \) for convenience. A flowchart of the bivariate uncertainty and correlation quantification process is shown in Figure 3.

Step 1. At the top of flowchart, data sets \( \{(X_j, Y_j)\}_{j=1}^n \) are collected in pairs from tests, where \( \{(X_j, Y_j)\}_j \) denotes the \( j \)-th observation data set and \( n \) is the number of data sets. Due to the small sample size, \( n \) is a small number in this paper. To construct the marginal distribution of variables, the data samples are divided to two part, i.e. \( \{X_j\}_{j=1}^n \) and \( \{Y_j\}_{j=1}^n \).

Step 2. Assuming the probability distribution type of the variables \( x \) is known according to historic data or experience, the interval of distribution parameters \( \theta = (\theta_1, \theta_2, \ldots, \theta_c)^T \) are estimated in bootstrap method or improved kernel density estimation in Ref. [35] based upon the samples \( \{X_j\}_{j=1}^n \). For example, if the variable \( x \) follows a normal distribution, then \( \theta = (\mu, \sigma^2)^T \). The lower and upper bounds of its interval estimation are \( \theta_{ij} = (\mu_{ij}, \sigma^2_{ij})^T \) and \( \theta_{ij} = (\mu_{ui}, \sigma^2_{ui})^T \) respectively. Similar process is also performed for the variable \( y \) as shown in the flowchart.

Step 3. After random sampling the parameters \( \theta \) and \( \lambda \) from their intervals, Hundreds of CDFs \( \{F_k(x)\}_{k=1}^s \) and \( \{G_k(y)\}_{k=1}^s \) can be plotted to construct the p-boxes representing the uncertainty of variables \( x \) and \( y \) respectively. Then Latin hypercube design is employed to uniformly recombine the CDFs instead of using all the combinations in order to reduce computational cost. As a result, \( s \) pairs of marginal distribution curves are obtained which is prepared for correlation parameter estimation.
Step 4. For the specified copula, the correlation parameter $\rho_k$ can be estimated in IFM method based on the sample pairs $\{(X_j, Y_j)\}_{j=1}^n$ and the $k$-th marginal distribution pairs $(F_k(x), G_k(y))$. As a result, copulas are constructed with each element of the parameter vector $\mathbf{\rho} = (\rho_1, \rho_2, \cdots, \rho_s)^T$ corresponding to each pair of marginal distribution curves in p-boxes.

Figure 3. A flowchart of the bivariate uncertainty and correlation quantification process

3.2. Uncertainty propagation

After the uncertainty quantification of input variables, uncertainty propagation is performed by PBA and nested sampling technique upon the joint distribution (constructed by copula function) in this paper. The flowchart is shown in Figure 4 and it contains four steps as bellow.

Figure 4. The flowchart of uncertainty propagation
Step 1. The copula functions \( \{C_k(u, v; \rho_k)\}_{k=1}^{s} \) with \( s \) correlation parameters are collected in uncertainty quantification process. Corresponding to each copula, the marginal distribution pair of variables \( x \) and \( y \) (\( F_k(x), G_k(y) \)) are also collected.

Step 2. The probability resample matrix \( \{[u^T, v^T]\}_{k} \) is generated by the copula function with correlation parameter \( \rho_k \), where \( u = [u_1, u_2, \ldots, u_N] \), \( v = [v_1, v_2, \ldots, v_N] \), \( N \) is the length of matrix and the number \( k \) starts from 1 to \( s \) in a loop. For every \( \rho_k \) in the loop, the elements of the matrix \( (u_j, v_j) \in [0,1]^2 \) are resampled from the copula function \( C_k(u, v; \rho_k) \in [0,1] \).

Step 3. The \( k \)-th resample pairs of input variables \( x \) and \( y \) \( \{[X_j, X_j]\}_{j=1}^{N} \) are generated by the probability resample matrix \( \{[u^T, v^T]\}_{k} \) and \( k \)-th pair of marginal distributions \( (F_k(x), G_k(y)) \). By calculating the model \( z = h(x, y) \) \( N \) times, the response resamples of interest are obtained as \( \{[Z_1, Z_2, \ldots, Z_N]\}_{k} \) and the probability distribution of the response \( R_k(z) \) can be represented with empirical CDF of the data because the sample size \( N \) is large enough.

Step 4. While \( k \) increases from 1 to \( s \), the p-box is constructed with a bunch of probability distributions \( R_k(z) \) to represent the uncertainty of response \( z \) of interest.

4. A numerical example
Assuming that both the correlated variables \( x \) and \( y \) follow normal distributions with unknown parameters, the uncertainty and correlation between them are quantified based on \( n \) pairs of samples in this Section. Supposing the response quantity \( z \) has a very simple relation of variable \( x \) and \( y \), i.e. \( z = x - y \), the uncertainty of \( z \) is also quantified in this Section.

It should be noted that the assumptions are drawn from a practical background and it describes a common situation widely happened in engineering. The uncertainty of numerous variables in engineering is demonstrated to follow normal distribution from the abundant existing data and the distribution parameters are usually determined with samples from physical tests or experiments. Moreover, the variables in the system are often correlated with each other as mentioned above. Furthermore, the correlation may be occurred between input and input variables, output and output variables, or input and output variables, e.g. the actual stress of structure and the yield stress of the material at key position. The reliability assessment of the structure may be obtained with the remainder subtracting actual stress from the yield stress.

To test the validity of the proposed method, sample sets are generated from the standard bivariate normal distribution with the correlation parameter \( \rho = 0.8 \). Uncertainty quantification and propagation at the situations of different simple sizes (i.e. 20, 100, and 1000) for input variables are accomplished. The scatter plot Figure 5 shows three kinds of samples generated by standard bivariate normal distribution with the correlation coefficient valued in 0.8.

![Figure 5](https://example.com/figure5.png)

**Figure 5.** The scatter plots of input variables' original samples generated by standard bivariate normal distribution with the correlation coefficient 0.8. The sample sizes in the plots are a) \( n=20 \), b) 100, c) 1000 respectively.

4.1. Uncertainty quantification and joint distribution construction of input variables
Since the probability distribution types of variables \( x \) and \( y \) are normal distribution, the mean \( \mu \) and variance \( \sigma^2 \) are estimated by Equation (7) from knowledge of mathematical statistics. The statistic quantities of the samples and parameter estimation results are all shown in Table 1. Then a bunch of
Gaussian copulas (the copulas’ number \( s=100 \)) are constructed and correlation parameters are estimated following the steps in Section 3.1. P-boxes of input variable \( x \) and the copula used to construct joint distribution of input variables are shown in Table 1 leaving out the p-boxes of variable \( y \) which look extremely similar as that of variable \( x \).

\[
\mu \in \left[ \bar{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\alpha/2} (n-1), \bar{X} + \frac{S}{\sqrt{n}} \cdot t_{1-\alpha/2} (n-1) \right]
\]

\[
\sigma^2 \in \left[ \frac{(n-1) \cdot S^2}{\chi^2_{1-\alpha/2} (n-1)}, \frac{(n-1) \cdot S^2}{\chi^2_{\alpha/2} (n-1)} \right]
\]

(7)

where \( \bar{X} \) and \( S^2 \) are the sample mean and sample variance respectively; \( \alpha_1 \) and \( \alpha_2 \) are the significant levels of the population mean and population variance. \( t_p (n-1) \) is the quantile of t-distribution with \( n-1 \) degrees of freedom under the probability \( p \).

**Table 1.** Statistics quantities of the samples and distribution parameters estimation of input variables

| Sample size | Sample corr. | Min. & Max. of corr. Estm. | Mean of sample | Interval Estm. of \( \mu \) | Variance of sample | Interval Estm. of \( \sigma^2 \) |
|-------------|--------------|---------------------------|---------------|---------------------------|-------------------|---------------------------|
| 20          | 0.786        | [0.564, 0.896]            | \( x \) 0.088 | -0.480 - 0.656            | 1.214             | 0.852 - 3.142             |
|             |              |                           | \( y \) 0.071 | -0.550 - 0.692            | 1.327             | 1.018 - 3.755             |
| 100         | 0.777        | [0.722, 0.821]            | \( x \) 0.009 | -0.181 - 0.199            | 0.957             | 0.706 - 1.236             |
|             |              |                           | \( y \) 0.006 | -0.197 - 0.209            | 1.023             | 0.807 - 1.412             |
| 1000        | 0.787        | [0.774, 0.798]            | \( x \) -0.026 | -0.086 - 0.034            | 0.963             | 0.852 - 1.015             |
|             |              |                           | \( y \) -0.039 | -0.100 - 0.021            | 0.978             | 0.878 - 1.046             |

**Figure 6.** The p-boxes of variable \( x \) constructed by 100 marginal distribution curves, which are based on the original samples with size of a) 20, b) 100, c) 1000, respectively. And a Gaussian copula function constructed by IFM method and original samples.
4.2. Uncertainty propagation

By the simple response model, i.e. \( z = x - y \), uncertainty propagation considering the correlation between input variables are performed and the p-box representing response uncertainty is drawn with red lines in Figure 7. Compared with that, p-box obtained by uncertainty propagation without considering the correlation between input variables is also drawn with blue lines in Figure 7. Only the margins of p-boxes instead of all curves in p-boxes are drawn to make the diagrams clear. Moreover, uncertainty propagation based on original samples with different sizes are all accomplished and drawn in Figure 7 a), b) and c) respectively.

![Figure 7](image)

**Figure 7.** The comparison between the p-boxes with considering and without considering the correlation between input variables. Only the margins of p-boxes instead of all curves in p-boxes are drawn to make the diagrams clear. The uncertainty propagation is based on data samples with size of a) 20, b) 100, and c) 1000 respectively.

In addition, both the nominal probability distributions of response considering the correlation between input variables and assuming independent variables are also drawn in to test the validity of p-boxes. As all the input variables follow normal distribution and the response has a linear relation with input variables, the response also follows a normal distribution from the statistics mathematic knowledge. Assuming the response \( z = ax + by \), the mean and variance of response can be express as:

\[
\text{mean} = a \mu_x + b \mu_y, \quad \text{variance} = a^2 \sigma_x^2 + b^2 \sigma_y^2 + 2ab \rho \sigma_x \sigma_y
\]

The results show that the p-box covers the theoretic distribution all the way whatever the sample size is, which confirms the validity of our uncertainty quantification method. It is also shown that the epistemic uncertainty cannot be conveniently ignored when samples are very limited just like the diagram in Figure 6 a). Moreover, compared between Figure 6 a), b) and c), the p-boxes of the input variables are narrowing as the size of original samples increases, which demonstrates that the epistemic uncertainty has reduced as more available test data or more information is appended.

The data in Table 1 also point to the same conclusion. The interval in the third column of Table 1 is bounded by the minimum and maximum of the correlation coefficients that we have estimated by IFM method. And this interval covers the correlation coefficient of the original samples all the time though it narrows again and again as the samples size diminishes in. The interval estimation of the distribution parameters, i.e. the mean and variance in normal distribution, behaves in same way as the correlation coefficient.
\begin{align}
E(z) &= a \mu_x + b \mu_y, \\
\text{Var}(z) &= a^2 \sigma_x^2 + b^2 \sigma_y^2 + 2ab \rho_{xy} \sigma_x \sigma_y,
\end{align}

where \( \mu_x \) and \( \mu_y \) are means of variables \( x \) and \( y \) respectively, \( \sigma_x \) and \( \sigma_y \) are the standard deviations of variables \( x \) and \( y \) respectively, \( \rho_{xy} \) denotes the correlation coefficient between variables \( x \) and \( y \). Hence the parameters of the two nominal distributions can obtained from Equation (8) with correlation coefficients valuing in 0 for independent inputs and 0.8 for correlated inputs respectively.

The results show that the nominal distribution always lies in the corresponding p-box without exceeding its bounds regardless of the correlation between input variables in Figure 7. In other words, the blue p-box covers the black dashed lines and the red p-box covers the magenta dotted lines. And as same as the uncertainty quantification results drawn in Figure 6, the p-box narrows too as more original samples are available in Figure 7. These issues confirm the validity of our uncertainty propagation method.

The results indicate again the epistemic uncertainty cannot be ignored when samples are very limited. Moreover, it is demonstrated that the response p-box with considering the correlation between input variables (marked with red lines in Figure 7) shows notable difference from the p-box without considering the correlation (marked with blue lines in Figure 7). And the difference seems quite distinct when the samples size is large as the comparison of Figure 7 a), b) and c). Consequently, the uncertainty propagation ignoring correlation among input variables may lead to inaccurate results especially for strongly correlated input variables. Furthermore, when the sample size is small the “band width”, representing the epistemic uncertainty from lacking in original samples, of p-box may be so great that the influence of correlation between input variables on response p-box seems a bit slight. But as the epistemic uncertainty is reduced by appending more and more test data, the p-box would narrow down and the influence of correlation would be more and more obvious.

5. Conclusion

For uncertainty quantification and propagation based on limited samples, it is important to address both the issues of epistemic uncertainty from lacking in samples and correlations between input variables. This paper proposes a copula-based method to construct a bunch of joint probability distributions and corresponding marginal distribution curves (p-box) for uncertainty quantification with considering correlations between input variables. Moreover, uncertainty propagation for correlated input variables proceeds by means of resampling from the results of uncertainty quantification. Then the numerical example is studied to validate the proposed methods.

The computational results of the numerical example indicate that the epistemic uncertainty is significant and cannot be conveniently ignored when available samples are very limited. And it is also indicated that correlations among input variables may significantly impact the uncertainty propagation results and the effect seems obvious when epistemic uncertainty is reduced.

References

[1] Oberkampf W L, Helton J C, Joslyn C A, et al. Challenge problems: Uncertainty in system response given uncertain parameters[J]. Reliability Engineering and System Safety. 2004, 85(1-3): 11-19.

[2] Hemez F M, Atamturktur S. The dangers of sparse sampling for the quantification of margin and uncertainty[J]. Reliability Engineering and System Safety. 2011, 96(9): 1220-1231.

[3] Helton J C. Quantification of margins and uncertainties: Conceptual and computational basis[J]. Reliability Engineering and System Safety. 2011, 96(9): 976-1013.

[4] Jaulin L, Kieffer M, Didrit O, et al. Applied interval analysis[M]. New York, NY: Springer-Verlag, 2001.

[5] Shafer G. Mathematical theory of evidence[M]. Princeton, NJ: Princeton University Press, 1976.

[6] Shah H, Hosder S, Winter T. A mixed uncertainty quantification approach using evidence
theory and stochastic expansions[J]. *International Journal for Uncertainty Quantification*. 2015, 5(1): 21-48.

[7] Zadeh L. Fuzzy sets as a basis for a theory of possibility[J]. *Fuzzy Sets and Systems*. 1978, 1: 3-28.

[8] De Cooman G. Possibility theory part I: Measure and integral theoretic groundwork; Part II: Conditional possibility; Part III: Possibilistic independence[J]. *International Journal of General Systems*. 1997, 25(4): 291-371.

[9] Park I, Amarchinta H K, Grandhi R V. A Bayesian approach for quantification of model uncertainty[J]. *Reliability Engineering and System Safety*. 2010, 95(7): 777-785.

[10] Cheung S H, Oliver T A, Prudencio E E, et al. Bayesian uncertainty analysis with applications to turbulence modeling[J]. *Reliability Engineering and System Safety*. 2011, 96(9): 1137-1149.

[11] Helton J C, Johnson J D. Quantification of margins and uncertainties: Alternative representations of epistemic uncertainty[J]. *Reliability Engineering and System Safety*. 2011, 96(9): 1034-1052.

[12] Sentz K, Ferson S. Probabilistic bounding analysis in the Quantification of Margins and Uncertainties[J]. *Reliability Engineering and System Safety*. 2011, 96(9): 1126-1136.

[13] Williamson R C, Downs T. Probabilistic arithmetic I: Numerical methods for calculating convolutions and dependency bounds[J]. *International Journal of Approximate Reasoning*. 1990, 4(2): 89-158.

[14] Ferson S, Kreinovich V, Ginzburg L, et al. Constructing probability boxes and Dempster-Shafer structures[R]. Albuquerque: Sandia National Laboratories, 2003.

[15] Helton J C, Johnson J D, Sallaberry C J, et al. Survey of sampling-based methods for uncertainty and sensitivity analysis[J]. *Reliability Engineering and System Safety*. 2006, 91(10-11): 1175-1209.

[16] Helton J C, Johnson J D, Oberkampf W L, et al. A sampling-based computational strategy for the representation of epistemic uncertainty in model predictions with evidence theory[J]. *Computer Methods in Applied Mechanics and Engineering*. 2007, 196(37-40): 3980-3998.

[17] Ferson S, Nelsen R B, Hajagos J, et al. Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis[R]. Albuquerque: Sandia National laboratories, 2004.

[18] Jiang C, Zhang W, Wang B, et al. Structural reliability analysis using a copula-function-based evidence theory model[J]. *Computers and Structures*. 2014, 143(Complete): 19-31.

[19] Jiang C, Wang B, Li Z R, et al. An evidence-theory model considering dependence among parameters and its application in structural reliability analysis[J]. *Engineering Structures*. 2013, 57(Complete): 12-22.

[20] Grzegorzewski P, Ziembinska P. Spearman’s rank correlation coefficient for vague preferences[J]. *Lecture Notes in Computer Science*. 2011, 7022(1): 354-363.

[21] Jiang C, Han X, Lua G Y, et al. Correlation analysis of non-probabilistic convex model and corresponding structural reliability technique[J]. *Computer Methods in Applied Mechanics and Engineering*. 2011, 200(33-36): 2528-2546.

[22] Sklar A. Fonctions de repartition a n dimensions et leurs marges[J]. *Publications del’Institut de Statistique del’Universite de Paris*. 1959, 8: 229-231.

[23] Sklar A. Random variables, joint distribution functions, and copulas[J]. *Kybernetika*. 1973, 9(6): 449-460.

[24] Nelsen R B. An introduction to copulas[M]. New York: Springer, 1999.

[25] Nelsen R B, Quesada-Molina J J, Rodriguez-Lallena J A, et al. Distribution functions of copulas: A class of bivariate probability integral transforms[J]. *Statistics and Probability Letters*. 2001, 54(3): 277-282.

[26] Genest C, Rivest L. Statistical inference procedures for bivariate Archimedean copulas[J]. *Journal of the American Statistical Association*. 1993, 88(423): 1034-1043.
Acknowledgements
The grant supports from National Natural Science Foundation of China (No. 11472256), the Science and Technology development Foundation of CAEP (No. 2013A0203007), key subject "Computational Solid Mechanics" of the CAEP, and the Presidential Foundation of CAEP (No. YZ2015011).