Reliability Engineering and System Safety 96 (2011) 117–130

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

A probabilistic approach for representation of interval uncertainty

Kais Zaman, Sirisha Rangavajhala, Mark P. McDonald, Sankaran Mahadevan *

Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 23 June 2009
Received in revised form
28 July 2010
Accepted 29 July 2010
Available online 4 August 2010

Keywords:
Interval uncertainty
Johnson distribution
Epistemic uncertainty

ABSTRACT

In this paper, we propose a probabilistic approach to represent interval data for input variables in reliability and uncertainty analysis problems, using flexible families of continuous Johnson distributions. Such a probabilistic representation of interval data facilitates a unified framework for handling aleatory and epistemic uncertainty. For fitting probability distributions, methods such as moment matching are commonly used in the literature. However, unlike point data where single estimates for the moments of data can be calculated, moments of interval data can only be computed in terms of upper and lower bounds. Finding bounds on the moments of interval data has been generally considered an NP-hard problem because it includes a search among the combinations of multiple values of the variables, including interval endpoints. In this paper, we present efficient algorithms based on continuous optimization to find the bounds on second and higher moments of interval data. With numerical examples, we show that the proposed bounding algorithms are scalable in polynomial time with respect to increasing number of intervals. Using the bounds on moments computed using the proposed approach, we fit a family of Johnson distributions to interval data. Furthermore, using an optimization approach based on percentiles, we find the bounding envelopes of the family of distributions, termed as a Johnson p-box. The idea of bounding envelopes for the family of Johnson distributions is analogous to the notion of empirical p-box in the literature. Several sets of interval data with different numbers of intervals and type of overlap are presented to demonstrate the proposed methods. As against the computationally expensive nested analysis that is typically required in the presence of interval variables, the proposed probabilistic representation enables inexpensive optimization-based strategies to estimate bounds on an output quantity of interest.

1. Introduction

As discussed in [1], there are three elements in model-based uncertainty analysis: (i) representing the input uncertainty, (ii) propagating the input uncertainty through a model of system response to obtain a representation of the output uncertainty and (iii) communicating the results thus obtained to the decision makers and other stakeholders. Therefore, it is important that all types of uncertainty in a model must be represented in a way that it can be efficiently used in further analysis, i.e., in algorithms for reliability analysis and design optimization and the results can be easily communicated to the stakeholders.

Two forms of uncertainty are commonly considered in the literature—aleatory and epistemic. Aleatory uncertainty is typically irreducible; examples include inherent variations in physical processes, such as weather conditions. Epistemic uncertainty represents a lack of knowledge about the system due to limited data, measurement limitations, or simplified approximations in modeling system behavior. This type of uncertainty can be typically reduced by gathering more information. Epistemic uncertainty can be viewed in two ways. It can be defined with reference to a stochastic but poorly known quantity [2] or with reference to a fixed but poorly known physical quantity [3]. The term "stochastic but poorly known" refers to the uncertainty about the distribution type and parameters of a *random variable*. It is now well recognized that both aleatory and epistemic uncertainty must be represented in an appropriate manner so that it can be used in any decision support analysis [1,4,5]. This paper focuses on handling the first definition of epistemic uncertainty, i.e., epistemic uncertainty with reference to a stochastic but poorly known quantity in a straightforward manner, as the uncertainty representation methods proposed in this paper are purely probabilistic, resulting in a family of probability distributions.

Within the context of reliability analysis, it is often required that a certain function $g(\mathbf{x})$ of input variables $\mathbf{x}$, representing a response of the designed system, lie within given bounds. In many cases, the values of some elements of $\mathbf{x}$ are uncertain, and this uncertainty may be of aleatory or epistemic type. Aleatory uncertainty can be represented by using probability distributions.

* Corresponding author.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

0951-8320/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2010.07.012

In some cases of epistemic uncertainty, the distribution for $\mathbf{x}$ must be determined from imprecisely available data, such as intervals given by experts. This implies that the cumulative distribution function of $\mathbf{x}$, and subsequently that of $g(\mathbf{x})$, denoted as $F(g(\mathbf{x}))$, cannot be known precisely. Instead of formulating design requirements in terms of failure probabilities, the requirements may then have to be formulated as bounds on the cumulative distribution function $F(g(\mathbf{x}))$ of the function $g(\mathbf{x})$. In this paper, we focus on the representation of epistemic uncertainty arising from interval data in the input variables $\mathbf{x}$, where a variable's possible values are described by intervals.

There are various probabilistic and non-probabilistic approaches for treating interval data, each with their own advantages and limitations as discussed in Section 2. In this paper, we propose a probabilistic approach to represent the uncertainty described by the interval data in order to take advantage of the fact that a probabilistic representation of all type of uncertainty enables the use of already well established probabilistic uncertainty propagation and design optimization methods. Further, a probabilistic uncertainty analysis results can be easily communicated to the decision makers as well as to other stakeholders.

The main contributions of this paper are summarized as follows. First, we present approaches based on continuous optimization to find the bounds on second and higher moments of interval data with single and multiple intervals (Section 3). Second, we demonstrate using numerical examples that these algorithms are scalable in polynomial time with respect to increasing number of intervals. Third, using the bounds on moments, we fit a family of Johnson distributions to interval data (Section 4). Analogous to the notion of empirical p-box as the bounding envelope for empirical distributions, we construct a Johnson p-box, which represents the bounding envelope for Johnson distributions for interval data. We discuss these contributions more elaborately in Section 2.

The remainder of the paper is organized as follows. Section 2 reviews the existing methods of the representation of epistemic uncertainty. Sections 3 and 4 outline the contributions of the paper as indicated above. Section 5 illustrates the proposed developments using different examples of interval data, where comparisons with alternate representations such as the empirical p-box are made. Section 6 concludes the paper with summary and future work.

2. Review of representation of epistemic uncertainty

2.1. Sources of interval data

Interval data are encountered frequently in practical engineering problems. Several such situations where interval data arise are discussed in [6,7], for example: (a) physical limits and theoretical constraints may be the only sources of information, which can only circumscribe possible ranges of quantities resulting in interval data. (b) Interval data arises when the only information available for a variable is a collection of expert opinions, which specify a range of possible values for a variable. (c) Reporting data with plus-or-minus uncertainties associated with the calibration of measuring devices also leads to interval data. (d) Some tests in chemical or purity quantification can only state that an observation is below a certain detection limit, resulting in an interval observation for the amount of impurity between zero and the threshold. (e) Intervals may arise in the detection of a fault when observations are spaced temporally; as the fault occurs between two consecutive observations, the time of failure is given by a window of time. Interval data requires careful treatment, especially if the width of the interval cannot be ignored when compared to the magnitude of the variable.

Two types of interval data are considered in this paper, based on computational methods: single or multiple intervals. When compared to single interval cases, multiple intervals require consideration of two additional issues: (1) from the context of computational expense, estimating statistics from multiple intervals can be more challenging (discussed later), (2) from the context of aggregation of information represented in the multiple intervals, there may be no basis to believe that the "true" value of the variable lies at any particular location of any intervals, such as endpoints or midpoints of the intervals. Although not necessarily true, a common assumption in the literature is that all the intervals are equally likely to enclose the "true" value of the variable, i.e., all intervals have an equal weight [6].

When data is available in multiple intervals (e.g., given by multiple experts), the information contained in one interval could contradict that in the other interval(s), or could be contained by other interval(s). In this context, intervals can be broadly categorized as *non-overlapping* and *overlapping* intervals.

2.2. What does an interval represent?

In order to propagate uncertainty through models of system response, it is necessary to first have a valid representation of the input uncertainty that can lead to meaningful quantification of the uncertainty in the system response. In this context, there are two broadly categorized interpretations of what interval data represents in the literature.

The first is the so-called equi-probability model, which corresponds to the Laplacian principle of indifference [8] and considers each interval as a uniform distribution [9]. Each possible value in every interval is assumed equally likely, resulting in a single probability mass and/or density for each possible realization of a random variable. We note that there might not be a justification to assume uniform distribution or any other distribution within a particular interval, which can be viewed as a limitation of the equi-probability model. Also, the equi-probability model results in a precise probabilistic representation of interval data, thereby failing to capture the inherent imprecision in the data.

The second popular interpretation of interval data, which is adopted in this paper, is that it represents *incertitude* in the data [6]. As a result, the possible values for quantities of interest such as probability of an event will in general be an *interval*, unlike a single value for point data. Unlike the equi-probability model that results in a single probabilistic representation of the interval, the notion of incertitude leads to a collection of distribution functions that could arise from different possible combinations of values from within the intervals.

The set of all possible probability distributions of a particular distribution type (e.g., empirical, normal) for a variable described by interval data is known as a probability box, or a p-box for short [10]. To illustrate, we explain the notion of an empirical p-box that exists in the literature [6], which is the collection of all possible empirical distributions for the given set of intervals. An empirical p-box summarizes the interval data set graphically. It is constructed as an increasing step function with a constant vertical step height of $1/N$, where N is the number of intervals. The construction of the empirical p-box requires sorting the lower and upper bounds for the set of intervals, followed by plotting the empirical cumulative distribution function (CDF) of each of the sorted bounds as shown in Fig. 1 for overlapping interval data with five intervals [5, 6; 5.5, 6.1; 6, 6.5; 5.4, 6.2; 5.6, 6.6] and non-overlapping interval data with four intervals [5, 6; 6.1, 6.7; 6.9,

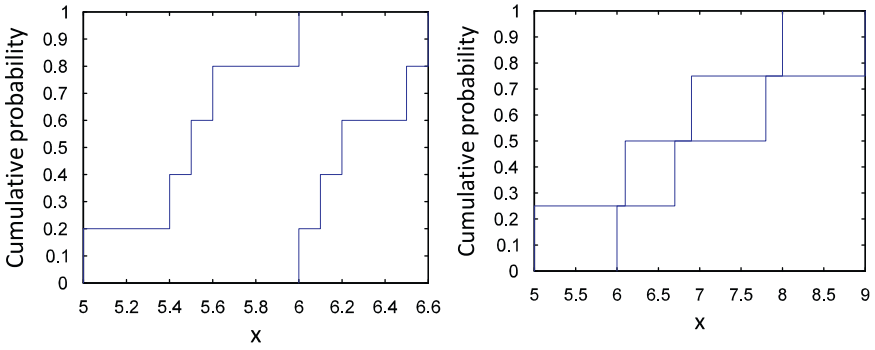

Fig. 1. Examples of an empirical p-box for multiple intervals. Left, overlapping data with 5 intervals [5, 6; 5.5, 6.1; 6, 6.5; 5.4, 6.2; 5.6, 6.6]; and right, non-overlapping data with 4 intervals [5, 6; 6.1, 6.7; 6.9, 7.8; 8, 9].

7.8; 8, 9]. The step height at each data point for the empirical CDF in Fig. 1 is equal, which reflects the assumption that the intervals are all equally weighed. Note that the p-box and the Dempster–Shafer structure (discussed in the following subsection) are equivalent and each representation can be converted to the other [11].

Several other approaches such as evidence theory and fuzzy logic are also used within the interpretation of incertitude. A brief discussion of the various techniques to represent interval data within the scope of incertitude is presented next.

2.3. Existing methods for treatment of interval uncertainty

The Sandia epistemic uncertainty project [12] conducted a workshop that invited various views on quantification and propagation of epistemic data uncertainty (includes interval data), which are summarized in [13]. Many uncertainty theories for representation and propagation of interval uncertainty have been discussed at the workshop, which include Dempster–Shafer structures [3,14], probability distributions [3], possibility distributions [3], subjective probabilities [15], random intervals [16], sets of probability measures [16], fuzzy sets [16], random sets [17,18], imprecise coherent probabilities [19], coherent lower previsions [20], p-boxes [21], families of polynomial chaos expansions [22], info-gap models [23], etc. A brief discussion of some of the popular uncertainty theories discussed in the above workshop, and interval data in general, follows.

In addition to the p-box representation discussed previously, other research within the realm of probability theory for interval data has focused on developing bounds, e.g., on CDFs. Hailperin [24] extensively developed the idea of interval bounds on CDFs as well as methods for propagation of these probability intervals through simple expressions. Hyman [25] developed similar ideas for probabilistic arithmetic expressions in the density domain. Williamson and Downs [10] described algorithms to compute arithmetic operations (addition, subtraction, multiplication and division) on pairs of p-boxes. These operations generalize the notion of convolution between probability distributions [26–28]. Additional results involving bounds on CDFs are available in [3,29].

Epistemic uncertainty has also been expressed using subjective probability (e.g., [15,30–31]). O'Hagan and Okaley [15] argue that probability is the only suitable representation of both aleatory and epistemic uncertainty. They also discuss the presence of imprecision in probability judgments, and argue that the problem is one of elicitation, not representation. This Bayesian approach [8] has been frequently used in the literature to include epistemic uncertainty, through the choice of prior distributions which are updated in the presence of new data. On the other hand, some researchers argue that a probabilistic representation is not appropriate for interval data because information may be added to the problem [7,32].

A commonly used approach for representation of interval data is Dempster–Shafer evidence theory [33]. In the context of evidence theory, there exist many rules to aggregate different sources of information. Among different rules of combination, the Dempster's rule is one of the most popular, however, this approach may not be suitable particularly for cases where there is inconsistency in the available evidence [32,34], e.g., in the case of non-overlapping intervals. In such cases, a mixture or averaging rule may be appropriate [34]. Evidence theory has been applied to quantify epistemic uncertainty in the presence of interval data for multidisciplinary systems design [32], where a belief measure is used to formulate the non-deterministic design constraints. Others have developed approaches for epistemic uncertainty quantification based on evidence theory, including [35,36].

In some cases, uncertain events form patterns that can be modeled using convex models of uncertainty [37]. Examples of convex models include intervals, ellipses or any convex sets. Convex models usually require less detailed information to characterize uncertainties than probabilistic models. They require a worst-case analysis in design applications which can be formulated as a constrained optimization problem. When the convex models are intervals, techniques in interval analysis can be used, though they are computationally expensive.

Some research has focused on use of possibility/fuzzy set theory for interval data. The possibility distribution (membership function) of a function of an interval variable with a given possibility distribution can be found using Zadeh's Extension Principle [38]. The drawback of this approach is that it requires combinatorial interval analysis, and the computational expense increases exponentially with the number of uncertain variables and with the nonlinearity of the function. Within the realm of fuzzy representation, Rao and Annamdas [39] present the idea of weighted fuzzy theory for intervals, where fuzzy set based representations of interval variables from evidences of different credibilities are combined to estimate the system margin of failure.

The aggregation of multiple sources of information as seen with multiple interval data is an important issue in characterizing input uncertainty. There is now an extensive lists of literature that discuss different aggregation methods, which include stochastic mixture modeling [3,21], Dempster's rule [32,40], a posteriori mixture [22], natural extension of pointwise maximum [20], etc. However, the aggregation method used in uncertainty representation must be consistent with the nature of the uncertainty as well as the specific uncertainty theory used [3].

Helton et al. [3] discussed and illustrated the use of different uncertainty theories, namely, probability theory, evidence theory, possibility theory, and interval analysis for the representation and propagation of epistemic uncertainty. This paper used a

sampling-based approach with each of the uncertainty theories. For probability theory, they defined the probability spaces by assuming uniform distributions over the sets of the possible values of the input variables. Multiple sources of information are aggregated by simply averaging the distributions for the number of sources assigning equal weight to each source. Baudrit and Dubois [2] proposed a methodology to represent imprecise probabilistic information described by intervals using different uncertainty approaches, such as probability theory, possibility theory, belief functions, etc.

Within the context of uncertainty propagation with interval variables, there exists literature that considers both interval and aleatory uncertainties. Approaches such as evidence theory or possibility theory are commonly used to represent interval variables, while probabilistic representation is typically used to represent aleatory uncertainties. The propagation of an evidence theory representation of uncertainty through a model of system response is computationally more expensive than that of probability theory [41]. Helton et al. [29] discussed the efficiency of different alternatives for the representation and propagation of epistemic uncertainty and argued that propagation of epistemic uncertainty using evidence theory and possibility theory required more computational effort than that of probability theory. In uncertainty propagation analysis, for every combination of interval values, the probabilistic analysis for aleatory variables is repeated, which results in a computationally expensive *nested* analysis. Some research in the literature focuses on managing this computational expense [42,43]. Representation and propagation of interval uncertainty has been studied from the context of structural problems [44] and multidisciplinary problems [45]. Besides their computational complexity, another disadvantage of using non-probabilistic methods is that the end users of the uncertainty analysis are a little aware of these methods and therefore, it may involve huge educational effort to make them familiar with these non-traditional uncertainty analysis methods [29].

2.4. Our contributions

As discussed above, there are various approaches for treating interval data, each with their own advantages and limitations. One of the drawbacks of the current approaches is the need for nested analysis in the presence of interval variables. To alleviate this issue, we propose a probabilistic representation for interval data using a family of Johnson distributions (see Appendix 1). A new aggregation technique is proposed to combine multiple intervals. This aggregation technique enables the use of the method of matching moments to represent the uncertainty described by the multiple intervals through a family of probability distributions. An important advantage of the proposed approach is that it allows for a unified probabilistic framework to be applied that can jointly handle aleatory and epistemic uncertainties, thereby allowing for well developed and efficient analytical probabilistic methods such as first-order reliability method (FORM) and second-order reliability method (SORM) [46] to be used in uncertainty propagation. The proposed representation avoids the expensive nested analysis by enabling the use of an optimization-based strategy that can estimate the distribution parameters of the input variables that minimize or maximize an output quantity of interest.

As mentioned earlier, probabilistic representation for interval data has attracted the criticism of assuming more information than what is available. This criticism is valid when a single distribution is forced on the data. However, by using a *family* of distributions to describe the interval data, as against a single distribution, we stay within the notion of incertitude inherently present in interval data discussed earlier. We note that a family of CDFs is possible for an interval random variable; however, it does not assert that any one CDF in the family is more or less likely to be the *true* CDF than the others.

It is a common practice in the literature to use methods such as moment matching and percentile matching to fit probability distributions to data sets. However, describing interval data in a probabilistic format is not straightforward. Unlike point data, where statistics such as moments have precise values, statistics for interval data are usually described by their upper and lower bounds. Note that by statistics for interval data, we mean the statistics of the distribution of the variable that is described by interval data. Finding bounds on the statistics of interval data is a computationally challenging problem because it typically involves interval analysis that is conducted using a combinatorial search. It has been reported that computing the upper bound on second moment of overlapping intervals is an NP-hard problem [6,47], although polynomial time algorithms have been reported for some special cases [48]. Little to no work exists in the literature about bounds on higher moments. Most previous approaches that calculate bounds on moments combinatorially search for points within the intervals that minimize or maximize the moments of the data. A major contribution of this paper is the development of algorithms based on continuous optimization methods which scale polynomially in computational effort with respect to the number of intervals. Knowledge of the bounds on moments on the interval data is useful because it provides restrictions on the possible distributions the underlying random variable may assume. Using the moment bounds computed using the proposed algorithms, we develop a probabilistic representation of the interval as a Johnson p-box, which is an ensemble of bounded Johnson distributions.

We note here that in this paper, the overall approach is similar to that of Ferson et al. [6], as this is the only work available in the current literature that discusses the possibility of estimating moment bounds for interval data and thus fitting a family of probability distributions to interval data. However, we are not dependent on any method developed or any assumption unique to Ferson et al. [6]. We have developed efficient algorithms for calculating moment bounds for interval data as opposed to computationally expensive heuristic approach used in [6] and unlike Ferson et al. [6], we have used a flexible family of distributions and also an optimization-based approach to represent interval data. This combination of techniques makes it feasible to include interval data in probabilistic analysis.

3. Estimating bounds on moments for interval data

This section discusses the proposed algorithms that estimate bounds on moments for interval data for single and multiple interval cases. A brief background is provided first.

In this paper, we fit a family of Johnson distributions to interval data using the moment matching approach. Moment matching involves equating the moments derived from data to those of the probability distribution being fit. The Johnson family is a generalized family of distributions that can represent normal, lognormal, bounded, or unbounded distributions. While there are several other viable four-parameter distributions that may also be used with this approach, such as the Pearson, Beta, and Lambda distributions, the Johnson family is a convenient choice. This is because the Johnson distribution lends itself to easy transformation to a standard normal space, which then can be conveniently applied in well known reliability analysis and reliability-based design optimization methods.

Among other methods (Appendix 1), we use the moment matching approach in this paper to take advantage of the moment bounding algorithms developed in this section. Moreover, to determine the appropriate type of Johnson distribution (bounded, unbounded, normal, lognormal), we need to compute the moments of the data set. While it is possible to have point estimates for the moments of point data, moments on interval data must be described using upper and lower bounds. As discussed in Section 1, it is challenging to compute bounds on moments of a variable described by multiple intervals. Note that in this paper, we assume that the multiple interval data are obtained from equally credible sources. As discussed in Section 2, this is a common assumption in the literature. The reason is that in absence of any additional information regarding the relative credibility of each source; it is reasonable to assume that all sources of information are equally credible.

In the following subsections, we propose methods that can compute lower and upper bounds on the first four moments for single and multiple interval cases.

3.1. Bounds on moments for single interval

In this subsection, we outline the proposed method to estimate bounds on moments for a single interval case.

In order to estimate the bounds on moments, we first find the probability mass function (PMF) of the endpoints of the interval that minimize or maximize the moments of the single interval data. The following procedure is used:

1. Sample ns data points from the given interval (both endpoints included).
2. Solve the following optimization problems with the PMFs, $p(x_i)$, $i=\{1, \ldots, ns\}$, as the decision variables:

$$\min_{p(x_i)}/\max M_k, \quad k = 1,2,3, \text{ or } 4 \tag{1}$$

such that

$$\sum_{i=1}^{ns} p(x_i) = 1 \tag{2}$$

Here, $M_1 = E(x)$

$$M_2 = E(x^2) - (E(x))^2$$
$$M_3 = E(x^3) - 3E(x^2)E(x) + 2(E(x))^3$$
$$M_4 = E(x^4) - 4E(x^3)E(x) + 6E(x^2)(E(x))^2 - 3(E(x))^4 \tag{3}$$

where $E(x) = \sum_{i=1}^{ns} x_i p(x_i)$

$$E(x^2) = \sum_{i=1}^{ns} x_i^2 p(x_i)$$
$$E(x^3) = \sum_{i=1}^{ns} x_i^3 p(x_i)$$
$$E(x^4) = \sum_{i=1}^{ns} x_i^4 p(x_i) \tag{4}$$

Note that the above formulas for the third and fourth moments have been derived from the definition of moments as given below [49].

Consider a random variable X for which the first moment, i.e., the expectation of X is $E(X)=\mu$. Then for any positive integer k, the expectation $E[(X-\mu)^k]$ is called the kth central moment of the variable X or the kth moment of X about the mean value.

3.1.1. Bounds on first moment for single interval

For the lower bound on the first moment, the above minimization yields that the probability mass function (PMF) at the lower endpoint of the interval is the Dirac delta function, i.e., PMF is equal to one at this point and zero elsewhere. Thus, the lower bound on the mean for a single interval is the lower bound of the interval itself. Similarly, the upper bound on the mean for a single interval occurs when the probability mass function (PMF) at the upper endpoint of the interval is the Dirac delta function. The upper bound on the mean for a single interval therefore is the upper bound of the interval. If we estimated the bounds on the first moment of single interval data based on observation, we would get the exact same results.

3.1.2. Bounds on second moment for single interval

For the lower bound on the second moment, the above minimization yields that the PMF at any point within the interval is the Dirac delta function, which implies that the lower bound on variance for a single interval is zero. Similarly, for the upper bound on the second moment, the above maximization yields a PMF of 0.5 at the both endpoints of the single interval.

3.1.3. Bounds on third moment for single interval

For the lower bound on the third moment, the above minimization yields a PMF of 0.2113 for the lower endpoint and 0.7887 for the upper endpoint. Similarly, a PMF of 0.7887 for the lower endpoint and 0.2113 for the upper endpoint is obtained for the upper bound on the third moment (maximization).

3.1.4. Bounds on fourth moment for single interval

For the lower bound on the fourth moment, the above minimization yields that the PMF at any point within the interval is the Dirac delta function, which implies that the lower bound on the fourth moment for a single interval is zero. For the upper bound on the fourth moment, the above optimization yields a PMF of 0.7887 for one of the endpoints and 0.2113 for the other.

We summarize these methods in Table 1. Note that these values of PMFs for the endpoints hold irrespective of the actual data represented by the single interval. For a given single interval, one could therefore directly use the above PMFs to estimate the lower and upper bounds on the moments, without having to repeat the optimization for each problem. We also note that we have solved the above mentioned optimization problems for four different sample sizes, i.e., by discretizing the single interval into four different sizes (10, 100, 500, and 1000) and obtained the exact same results with linear computational efforts. The nature of the sampling or discretization does not have any effect on the end results as long as the samples include the two endpoints of the single interval data.

It is seen from the optimization results that the minimum and maximum of the moments occur, when all the probability masses are concentrated at the two endpoints only, with two exceptions as seen for the lower bounds on the second and fourth moments. This is intuitive for the lower and upper bounds on the first moment. However, for the other cases, we investigate this issue as follows.

With regard to the proposed algorithm, the following can be stated from the definition of moments as mentioned earlier in this section:

1. As the second and fourth moments are by definition positive, the lower bounds on these moments are zero with the probability mass function (PMF) being the Dirac delta function at any point within the interval.
2. As the moments are by definition, the expectation of powers of deviation from the mean value, these expectations are

Table 1
Methods for calculating moment bounds for single interval data.

Moment	Condition		Formula
	Lower bound	Upper bound	
1	PMF=1 at lower endpoint =0 elsewhere	PMF=1 at upper endpoint =0 elsewhere	$M_1=E(x)$
2	PMF=1 at any point =0 elsewhere	PMF=0.5 at each endpoint	$M_2=E(x^2)-(E(x))^2$
3	PMF=0.2113 at lower endpoint =0.7887 at upper endpoint	PMF=0.7887 at lower endpoint =0.2113 at upper endpoint	$M_3=E(x^3)-3E(x^2)E(x)+2(E(x))^3$
4	PMF=1 at any point =0 elsewhere	PMF=0.7887 at one of the endpoints =0.2113 at the other endpoint	$M_4=E(x^4)-4E(x^3)E(x)+6E(x^2)(E(x))^2-3(E(x))^4$

Note: $E(x)=\sum_{i=1}^{2} x_i p(x_i) E(x^2)=\sum_{i=1}^{2} x_i^2 p(x_i) E(x^3)=\sum_{i=1}^{2} x_i^3 p(x_i) E(x^4)=\sum_{i=1}^{2} x_i^4 p(x_i)$ where $p(x_i)$=probability mass function (PMF).

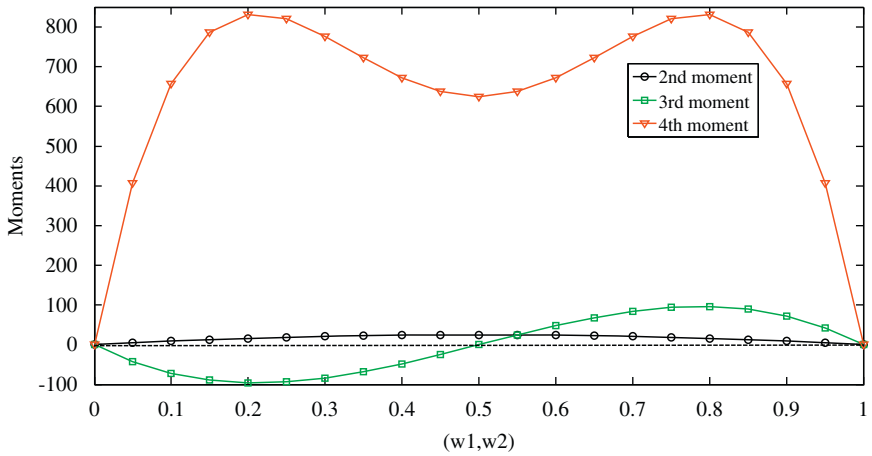

Fig. 2. Moments vs. PMFs at the interval endpoints.

essentially minimum (for the third moment) or maximum (for the second, third and fourth moments), when the data points are located at the endpoints of the interval, i.e., when the PMFs are concentrated only at the endpoints of the single interval.

Once we know that for minimum and maximum of some moments, the PMFs concentrate only on the two endpoints of the single interval, it might be interesting to investigate the nature of the solutions. We plot the values of the moments as a function of the pair (w_1, w_2), where w_1 is the PMF at the lower endpoint and $w_2=1-w_1$ is the PMF at the upper endpoint of the interval. It is seen in Fig. 2 that the second moment reaches its maximum when PMFs at both the endpoints are 0.5 each, which are consistent with our optimization results. For the third moment, we get a symmetric shape, which is consistent with our optimization results, where we have found that the PMFs at both the endpoints get flipped for the minimization (0.2113, 0.7887) and maximization problems (0.7887, 0.2113). For the fourth moment, we get a bi-modal shape. The curve reaches its maximum for two sets of PMF pairs (0.2113, 0.7887) and (0.7887, 0.2113), which are consistent with our optimization results. These two sets of PMFs also correspond to the minimum and maximum of the third moments, respectively, as seen in Fig. 2.

3.1.5. Numerical example

We apply the proposed method of estimating bounds on a single interval to the following example: [5,15]. The bounds on the first moment are calculated to be [5,15], those on the second moment are [0, 25], those on the third moment are [−96.225, 96.225], and those on the fourth moment are [0, 833.333]. We use this example later in the paper to illustrate subsequent steps in the proposed methodology.

3.2. Bounds on moments for multiple intervals

As discussed in Section 1, the computation of bounds on moments for multiple intervals is computationally expensive as it is usually treated as a combinatorial problem, where the moments are calculated at the combinations of possible values of the interval variable. Rather than deal with this problem combinatorially, we have formulated this computation as a nonlinear programming problem with the objective being minimization or maximization of the moments of data points that are constrained to fall within each of the respective intervals. The computational effort of this approach with increasing number of variables is demonstrated to be of polynomial order in the number of intervals. The proposed formulations are valid for any type of interval data, i.e. overlapping or non-overlapping intervals. The bounds on moments thus found are rigorous, i.e., they completely enclose all possible moments generated from various combinations of the interval data.

3.2.1. Bounds on first moment for multiple intervals

Consider a set of intervals given as $a_i \le x_i \le b_i$, $i=\{1, \ldots, n\}$, where n is the number of intervals. Estimating the bounds on the first moment (arithmetic mean) involves identifying a configuration of scalar points $\{x_i, i=\{1, \ldots, n\}\}$, (where x_i indicates the true value of the observation within the interval) within the respective intervals that yield the smallest possible mean, and a configuration that yield the largest possible mean. Because the

mean is proportional to the sum of the interval data, the configuration for the lower bound on the mean is the set of left endpoints of the interval, and that for the upper bound on the mean is the set of right interval endpoints. The formula for the arithmetic mean of interval data x_i is therefore

$$[\underline{M},\overline{M}] = \left[\frac{1}{n}\sum_{i}^{n} a_i, \quad \frac{1}{n}\sum_{i}^{n} b_i\right] \tag{5}$$

where $[\underline{M},\overline{M}]$ are the lower and upper bounds on the mean, respectively.

3.2.2. Bounds on second moment for multiple intervals

The second central moment (variance) is a quadratic function of each of the values of its data. We search for the configuration of scalar points, x_i, constrained to lie within their respective intervals that minimizes (or maximizes) the function shown below to yield the lower (or upper) bound on the variance. Therefore, we construct a linearly constrained optimization problem as follows:

$$\min_{x_1,\ldots,x_n}/\max_{x_1,\ldots,x_n} M_2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{6}$$

such that

$$a_i \le x_i \le b_i, \quad i = \{1,\ldots,n\} \tag{7}$$

3.2.3. Bounds on third and fourth moments for multiple intervals

The third and fourth central moments are third and fourth order polynomial functions of each of the values of the data, respectively. We search for the configuration of points $\{x_i,\ i=\{1, 2,\ldots,\ n\}\}$ constrained to lie within their respective intervals that minimizes (or maximizes) the function shown below to yield the lower (or upper) bound of the third/fourth moment.

$$\min_{x_1,\ldots,x_n}/\max_{x_1,\ldots,x_n} M_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^k \tag{8}$$

such that

$$a_i \le x_i \le b_i, \quad i = \{1,\ldots,n\} \tag{9}$$

where minimizing (or maximizing) the above problem with $k=3$ and 4 yields the lower (or upper) bound on the third and fourth moments, respectively.

We have implemented the formulations to calculate the lower and upper bounds on the second, third and fourth moments for various test cases with increasing number of intervals. We considered both overlapping and non-overlapping interval examples to demonstrate the performance of the proposed formulations. The following procedure was used to generate the intervals for overlapping interval test cases. The interval extremes (lowest of the lower bound and the highest of the upper bound) were arbitrarily assumed. In order to generate a desired number of intervals for each test case, a uniform random number generator was used to generate overlapping intervals between interval extremes. To generate non-overlapping interval data with n intervals for the test problems, we used the following procedure. First, a sequence of monotonically increasing random numbers is generated, $\{1\ ,\ldots,\ 2\times n\}$. The ith interval is generated by collecting the $(2i-1)$th and $(2i)$th random number. Thus the interval widths and the endpoints are generated randomly.

We solved the above optimization formulations in Eqs. (6)–(9) using the MATLAB function *fmincon*, which implements a sequential quadratic programming algorithm. The plots in the Figs. 3 and 4 illustrate the scalability of the proposed formulations with increasing number of intervals for overlapping and non-overlapping cases, respectively. For each plot shown in Figs. 3 and 4, we fit a linear or quadratic function as well as an exponential function (solely for comparison purposes). The regression coefficients (i.e., the values of R^2) indicate a strong linear/quadratic trend for the scalability of the algorithms.

Observe that the computational effort for estimating the lower bound on second moment increases linearly with increasing number of intervals for both overlapping and non-overlapping data (subplots (a) in both Figs. 3 and 4). The computational effort to estimate the upper bound on second moment with increasing number of intervals is observed to be $O(n^2)$, making this a computationally affordable procedure, even for relatively large data sets(subplots (b) in both Figs. 3 and 4).

The computational effort is also found to scale polynomially with the number of intervals for both minimization and maximization of third and fourth moments, as seen from subplots (c)–(f) in both Figs. 3 and 4. These plots show the best fitting polynomial and exponential trend lines to show that the trend is indeed polynomial in the number of intervals.

So far, we discussed the proposed optimization formulations to estimate bounds on the second, third, and fourth moments of interval data, which is the first important contribution of this paper. The moment bounds estimated in this section can be used to fit a family of Johnson distributions to interval data, as discussed in the next section.

4. Fitting Johnson distributions to interval data

As discussed in Appendix 1, there are several approaches to fit Johnson distributions to point data using statistics such as moments or percentiles. Unlike for point data where there can be a single probability distribution as the uncertainty description (when a large amount of samples is available), multiple probability distributions could describe interval data. Once the bounds on the moments of the interval data are calculated using the approach outlined in the previous section, we can now fit the Johnson distributions whose moments fall within the bounds of the moments of the interval data.

Within the proposed framework, two procedures could be adopted for the uncertainty quantification of interval data: (1) *sampling-based*, which involves taking random samples of moments from within the bounds computed earlier, and fitting a Johnson distribution to each set of sampled moments and (2) *optimization-based*, where a bounding envelope of the family of distributions can be constructed using an optimization approach using percentiles. The sampling based approach is discussed next.

4.1. Sampling-based procedure

The proposed sampling-based procedure for constructing the family of Johnson distributions is as follows:

1. Calculate the bounds on the first four moments of single or multiple interval data (Section 2).
2. Randomly select a set of moments from within the bounds of the first four moments. This sampling can be done using uniform distribution or by any discretization method. In this paper, we use uniform distribution. We note here that the type of sampling or discretization method used might have impact on the end results. However, this issue is not investigated in this paper.
3. From Fig. A1 (see Appendix), infer the type of distribution to be fitted (e.g. bounded, unbounded, etc.). We only select those samples that suggest a bounded Johnson distribution fit, so that the resulting distribution lies within the bounds of the

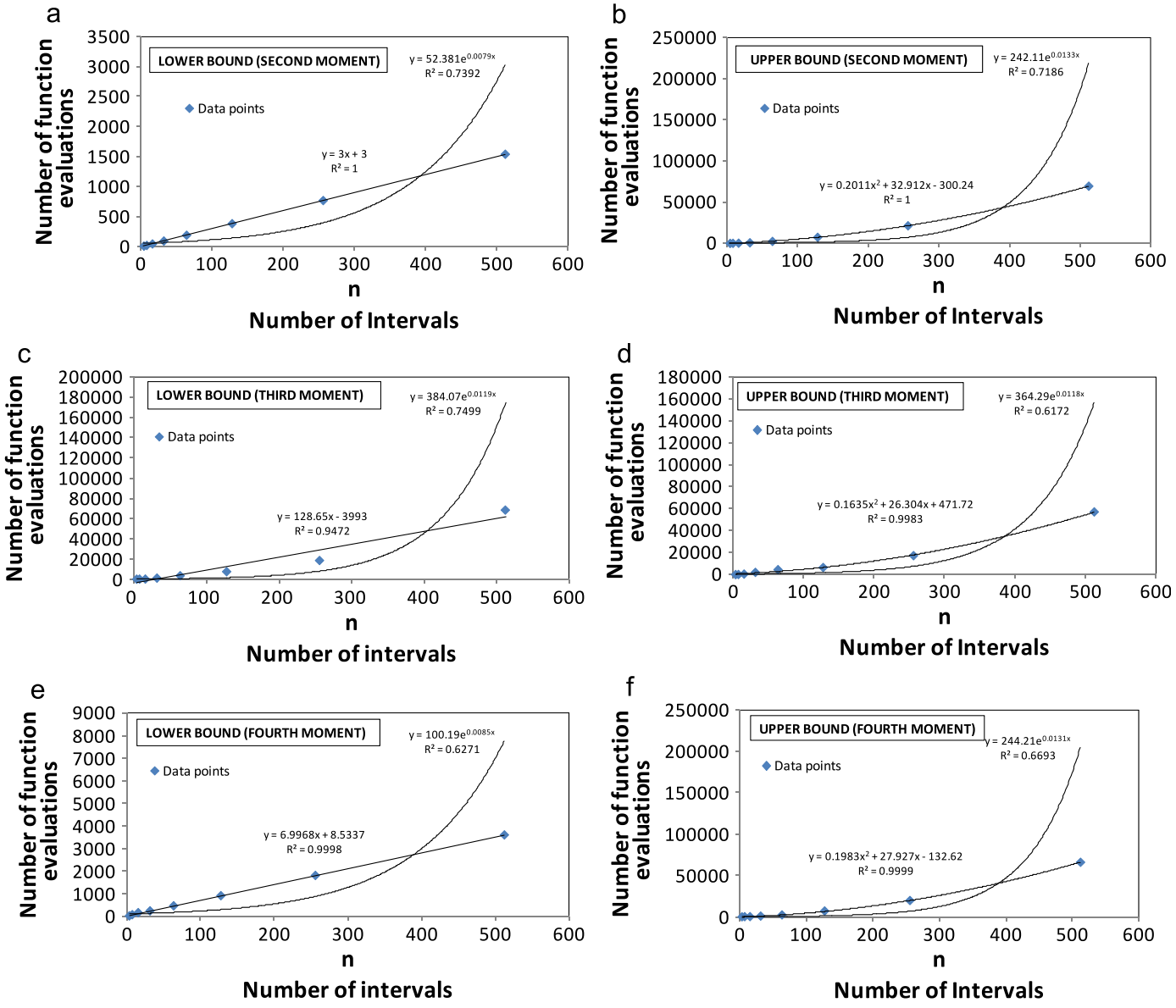

Fig. 3. Computational effort for the estimation of bounds on second, third, and fourth moments for overlapping intervals.

interval data specified, because, for interval uncertainty, it may be reasonable to argue that the true measurement has zero probability of lying outside the given interval for the single interval case or outside the overall bounds ([*Min*(*Lower bounds*) *Max*(*Upper bounds*)]) for the multiple interval case.

4. Using the bounds of the interval data, two parameters of the bounded Johnson distribution, ξ and λ, are estimated as $\xi = \min\{a_i,\ i=\{1,\ldots,\ n\}\}$, and $\lambda = \max\{b_i,\ i=\{1,\ \ldots,\ n\}\} - \min\{a_i,\ i=\{1\ ,\ldots,\ n\}\}$. The parameters ξ and λ, which are the location parameters [49], determine the lower end point and the range, respectively, of the bounded Johnson distribution.
5. The remaining two unknown parameters γ and δ, which govern the shape of the bounded Johnson distribution, are computed by solving the following optimization problem:

$$\min_{\gamma,\delta} \quad f(x) = \sum_{i=1}^{4} (M_{i(sampled)} - M_{i(johnson)})^2 \tag{10}$$

such that

$$-50 \le \gamma \le 50 \tag{11}$$

$$0.2 \le \delta \tag{12}$$

where $M_{i(sampled)}$ is the set of moments sampled from step 2, and are the set of moments for a $M_{i(johnson)}$ Johnson distribution. Constraints on the Johnson parameters are imposed for numerical reasons (discussed later). Note that the objective function of the above optimization problem may require scaling since the moments can be of largely different magnitudes.

6. Repeat steps 2–4 for a desired number of times. Each repetition of steps 3–5 yields a single Johnson distribution.

The above sampling-based procedure can be repeated as many times as desired to obtain a family of Johnson distributions. The issue of sampling size can be problem dependent. The sampling-based procedure of uncertainty representation cannot guarantee rigorous bounds on input distributions, as it might underestimate the uncertainty due to practical limitations or computational expense. The sample size is a more critical issue when this uncertainty has to be propagated through some models of system

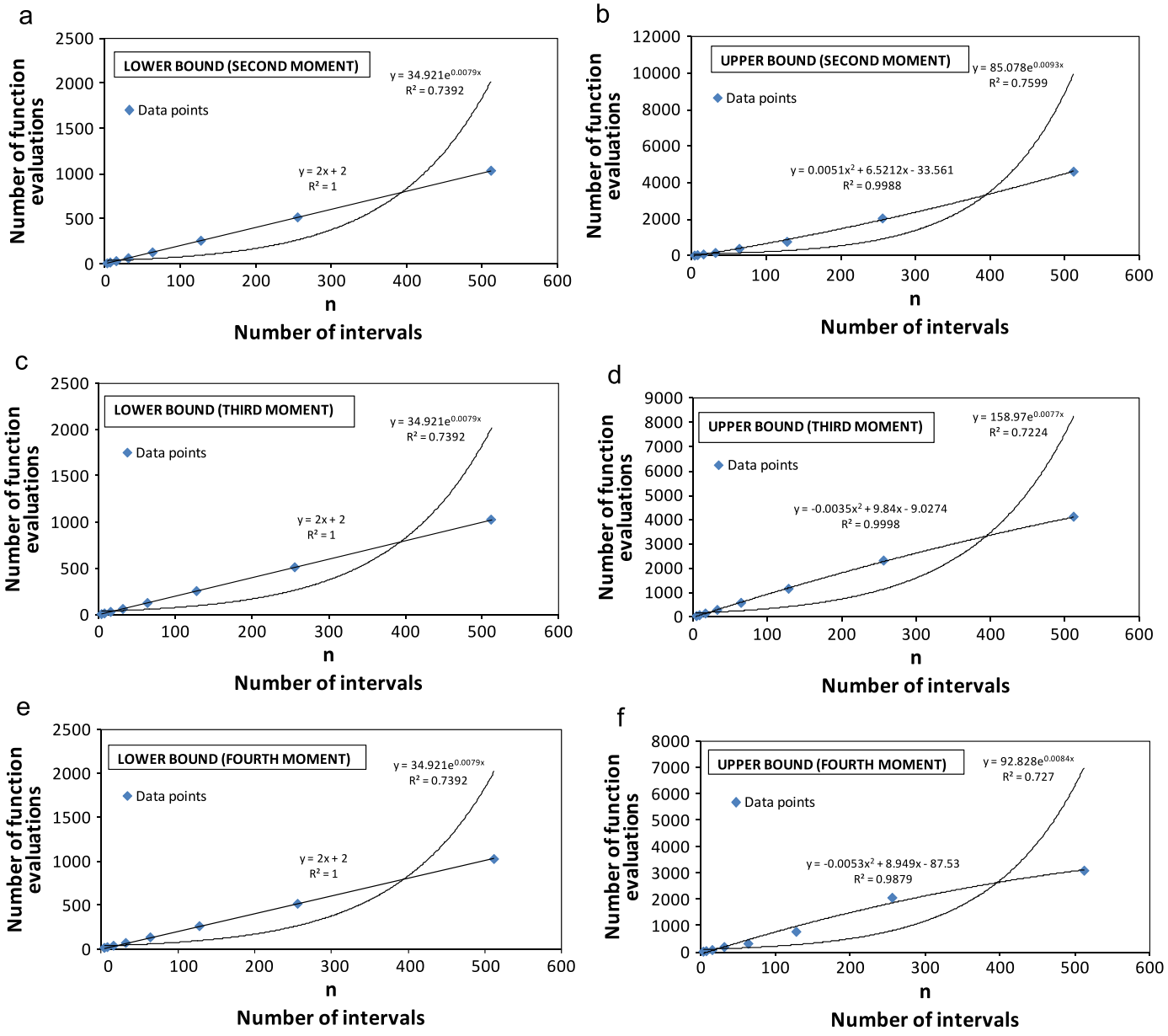

Fig. 4. Computational effort for the estimation of bounds on second, third, and fourth moments for non-overlapping intervals.

response. In order to alleviate the issue of sampling size in uncertainty representation, we have proposed an optimization-based strategy to represent interval uncertainty.

Note that the above procedure is the same for both overlapping and non-overlapping intervals. The optimization-based procedure to generate a probabilistic representation for interval data is discussed next.

4.2. Optimization-based procedure: Johnson p-box

Theoretically, infinitely many distributions can be fit to the given interval data. It is of interest for practical reasons to compute bounding envelopes for the family of Johnson distributions, which we call the Johnson p-box. Note that the Johnson p-box is analogous to the empirical p-box (Fig. 1), which is the bounding envelope of empirical distributions to fit the interval data. In this subsection, we present an optimization formulation based on percentiles to construct the Johnson p-box.

In order to compute the bounding envelope, we solve a set of optimization problems, each for a different percentile value, α, where $0.01 \leq \alpha \leq 0.99$. Each optimization problem for a chosen α finds the parameters of the Johnson distribution that maximize or minimize the Johnson variable, x^{α}, such that the moments of the Johnson distribution fall within the bounds computed in Section 2. The following optimization formulation is used to compute the Johnson p-box. Note that the minimization yields the left most bound of the family of distributions for each α. Similarly, maximization of the optimization problem below yields the right most bound for each α.

$$\min_{\gamma,\delta} / \max\ x^{\alpha} \tag{13}$$

such that

$$m1_{lb} \leq m1_{johnson} \leq m1_{ub} \tag{14}$$

$$m2_{lb} \leq m2_{johnson} \leq m2_{ub} \tag{15}$$

$$m3_{lb} \le m3_{johnson} \le m3_{ub} \quad (16)$$

$$m4_{lb} \le m4_{johnson} \le m4_{ub} \quad (17)$$

$$-50 \le \gamma \le 50 \quad (18)$$

$$0.2 \le \delta \quad (19)$$

where x^{α} is the αth percentile point, $0.01 \le \alpha \le 0.99$, $m1_{johnson}$, …,$m4_{johnson}$ are the first four moments of the Johnson distribution with parameters ξ, λ, γ, and δ, respectively, which can be computed using simulation; $m1_{lb}$, …, $m4_{lb}$, respectively, are the lower bounds on the first four moments of the interval computed using the proposed approach; and $m1_{ub}$, …, $m4_{ub}$, respectively, are the upper bounds on the first four moments of the interval computed using the proposed approach.

The value of the objective function, x^{α}, can be found by applying the Johnson transformation (see Eq. (A1) in the appendix) to a standard normal variable corresponding to the given α. Constraints in Eqs. (18) and (19) are imposed on the Johnson parameters for numerical reasons. The bounded Johnson transformation [50] is given as $x=\xi+\lambda[1+\exp(-((z-\gamma)/\delta)]^{-1}$, where x is the Johnson variable, and z is the standard normal variable. The δ parameter is restricted to be greater than 0.2: as $\delta \to 0$, the moments approach the impossible region for the Johnson family of distributions (Fig. A1) and can cause division by zero problems with the bounded Johnson transformation (Eqs. (A1) and (A2)). The bounds on γ have been chosen so that the bounded Johnson transformation function, $[1+\exp(-((z-\gamma)/\delta)]^{-1}$, has a finite non-zero value.

5. Numerical examples

In this section, we apply the proposed approaches to five example problems. We consider four multiple interval examples, each with different numbers of intervals and overlaps, and one single interval example. Comparisons with alternate representations, such as the empirical p-box, are also discussed.

5.1. Illustration of the proposed methodology

We consider two examples each for overlapping and non-overlapping multiple interval data, each with different numbers of intervals (Table 2). We follow the procedure outlined in Section 4.1 to fit a family of bounded Johnson distributions to each multiple interval data set in Table 2. The cumulative distribution functions of the family of Johnson distributions for each multiple interval data set are shown by thin dotted lines in Fig. 5. The corresponding single interval results, where the moment bounds are computed using the methods outlined in Section 2.1, are shown in the left hand side plot in Fig. 6.

The Johnson p-box optimization problem is solved for each set of interval data in Table 2 using Matlab's *fmincon* solver. We use 20 equally spaced points for the percentile values, α, ranging between 0.01 and 0.99. Note that the selection of the number of percentile points is arbitray. However, solving the optimization problem at increased number of percentile points results in more accurate bounds on uncertainty but with increased computational efforts. For each α, the minimization and maximization problems yield the left and right bounds on the p-box in Fig. 5, respectively. At each α value, we repeated the maximization/minimization using 15 different starting points to avoid local optima; the best results among the 15 runs are reported.

It is interesting to note that the Johnson p-boxes in Fig. 5 for all the multiple interval examples shown have discontinuities. It is noted that the set of active constraints in the optimization (particularly, those with the moment bounds (Eqs. (14)–(17)) changes at the point of discontinuity. For example, at the point A for Example 1 in Fig. 5, the set of active constraints changes.

Below the point A, the constraints on the upper bound of the third moment (upper bound in Eq. (16)) and on the lower bound of the first moment (lower bound in Eq. (14)) are active. Above the point A, the constraints on the lower bound of the fourth moment (lower bound in Eq. (17)) and on the lower bound of the first moment (lower bound on Eq. (14)) are active. Similar trend was observed at point B in Example 1, where a discontinuity occurs in the bounding envelope. Below the point B, the lower bound of the fourth moment (lower bound in Eq. (17)) and the upper bound of the first moment (upper bound in Eq. (14)) are the active constraints. Above the point B, the lower bound of the third moment (lower bound in Eq. (16)) and the upper bound of the first moment (upper bound in Eq. (14)) are the active constraints. For the single interval example, the Johnson p-box coincides with the left and right endpoints of the interval data.

5.2. Comparison with other representations

In this subsection, we present a comparison of the Johnson p-box with the empirical p-box idea available in the literature. We also compare how the choice of Johnson family of distributions impacts the probabilistic representation of interval data. Using an optimization formulation similar to that of the Johnson p-box, we compute the corresponding bounding envelopes for normal and lognormal distributions for single and multiple interval examples.

The empirical p-boxes for the multiple interval data cases, obtained by sorting the endpoints of the intervals, are also plotted in Figs. 5 and 7 for comparison purposes (thin solid lines). Note that for all examples presented in this section, not all members of the Johnson family of distributions fall inside the empirical p-box. The moments of the family of Johnson distributions fall within the moment bounds computed earlier; however, the distributions do not necessarily fall within the empirical p-box.

In order to study the effect of the choice of Johnson family, we compare the Johnson p-box to the bounding envelopes obtained for normal and lognormal distributions. The following optimization formulation is used to find the bounding envelopes for

Table 2
Interval data for the five numerical examples.

Example	Data
Example 1 with 5 overlapping intervals	[5, 6; 5.5, 6.1; 6, 6.5; 5.4, 6.2; 5.6, 6.6]
Example 2 with 9 overlapping intervals [6]	[3.5, 6.4; 6.9, 8.8; 6.1, 8.4; 2.8, 6.7; 3.5, 9.7; 6.5, 9.9; 0.15, 3.8; 4.5, 4.9; 7.1, 7.9]
Example 3 with 4 non-overlapping intervals	[5, 6; 6.1, 6.7; 6.9, 7.8; 8, 9]
Example 4 with 6 non-overlapping intervals [6]	[1, 1.52; 2.68, 2.98; 7.52, 7.67; 7.73, 8.35; 9.44, 9.99;3.66, 4.58]
Example 5 with a single interval	[5,15]

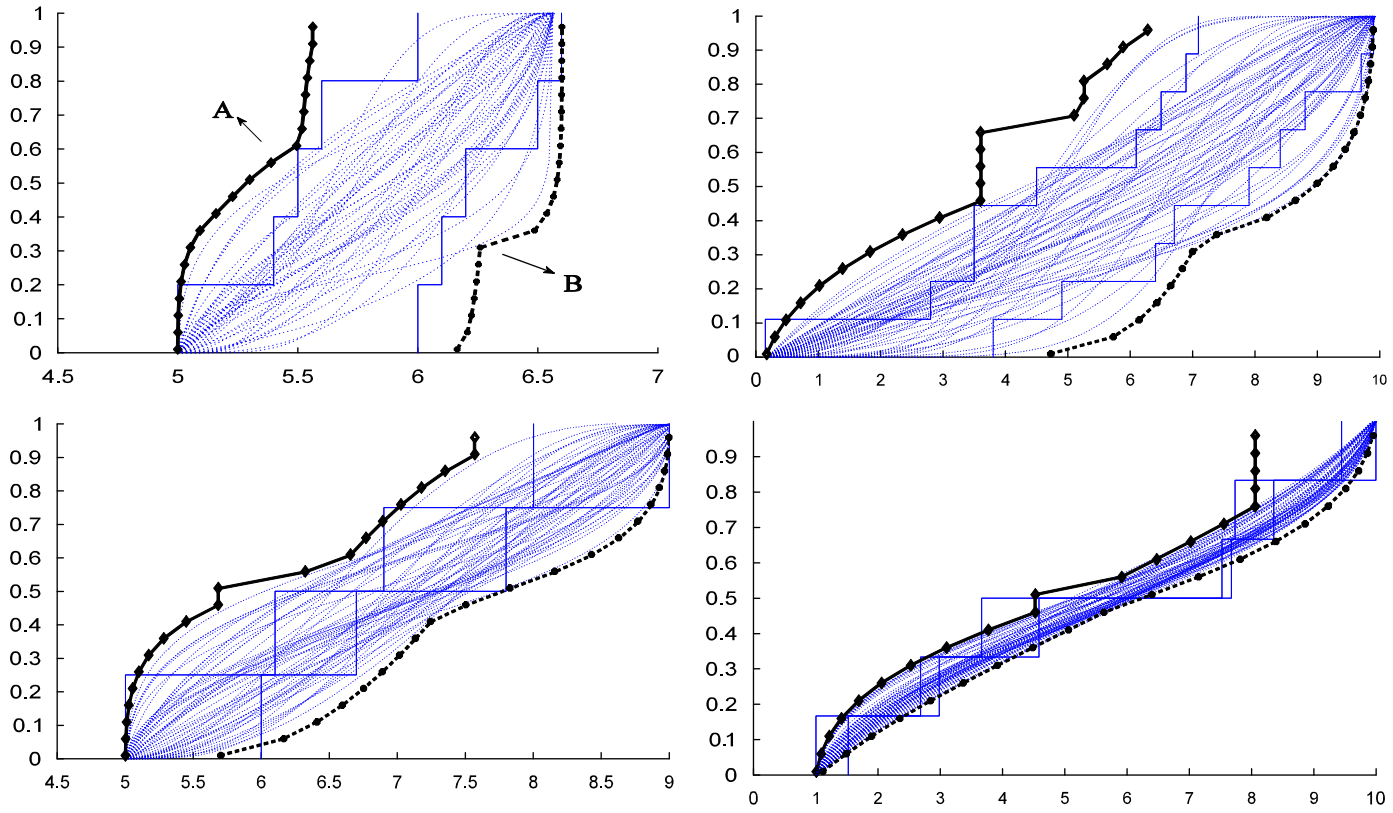

Fig. 5. Samples from the family of Johnson cumulative distributions for overlapping and non-overlapping examples for multiple interval examples (thick solid lines—Johnson p-box, thin solid lines—empirical p-box, and dashed thin lines—family of Johnson CDFs). Top left, Example 1; top right, Example 2; bottom left, Example 3; bottom right, Example 4.

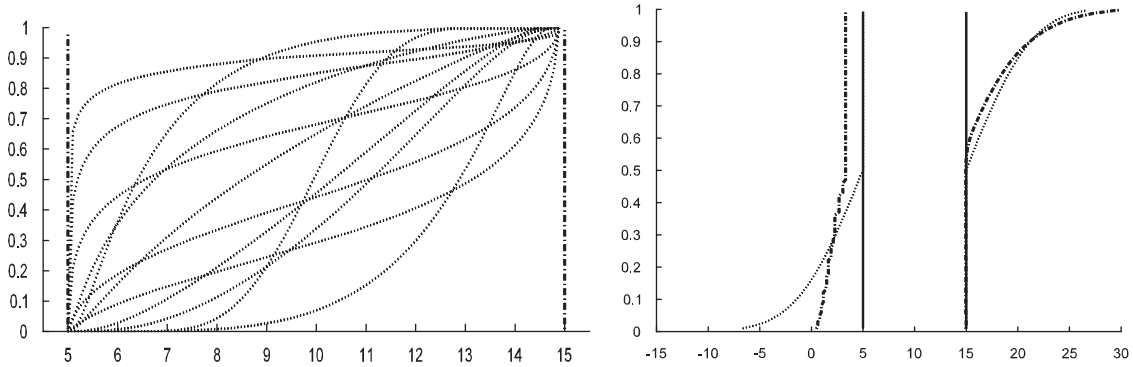

Fig. 6. Single interval example. Left, samples from the family of Johnson distributions (thin dotted lines) with Johnson p-box (thick dotted line); and right, comparison of Johnson p-box (thick solid line) with normal (thin dotted line) and lognormal (thick dotted line) p-boxes.

normal and lognormal distributions, where constraints are imposed on the first two moments.

$$\min_{d}/\max x^{\alpha} \tag{20}$$

such that

$$m1_{lb} \le m1_{dist} \le m1_{ub} \tag{21}$$

$$m2_{lb} \le m2_{dist} \le m2_{ub} \tag{22}$$

where x^{α} is the αth percentile point, $0.01 \le \alpha \le 0.99$, $d=(\mu_Y, \sigma_Y)$ is the design variable vector, where Y is the normal random variable; $m1_{dist}$ and $m2_{dist}$ are the first and the second moments for normal/lognormal distributions, respectively; $m1_{lb}$ and $m2_{lb}$, respectively, are the lower bounds on the first two moments of the interval computed using the proposed approach; and $m1_{ub}$ and $m2_{ub}$, respectively, are the upper bounds on the first two moments of the intervals computed using the proposed approach.

The quantities $m1_{dist}$ and $m2_{dist}$ for the normal p-box are related to the design variable vector, $m1_Y=\mu_Y$ and $m2_Y=\sigma_Y^2$. The moments of the lognormal variable (X), $m1_{dist}$ and $m2_{dist}$, are computed in terms of the corresponding normal variable moments, (μ_Y, σ_Y), as follows:

$$m1_X = e^{\mu_Y + 0.5\sigma_Y^2} \tag{23}$$

$$m2_X = \mu_X^2(e^{\sigma_Y^2}-1) \tag{24}$$

The maximization and minimization at each percentile point for the normal and lognormal have been repeated with 15 different starting points to avoid local optima. The best results from within the 15 starting points have been plotted in Fig. 7. Note that the bounded Johnson p-box remains close to the

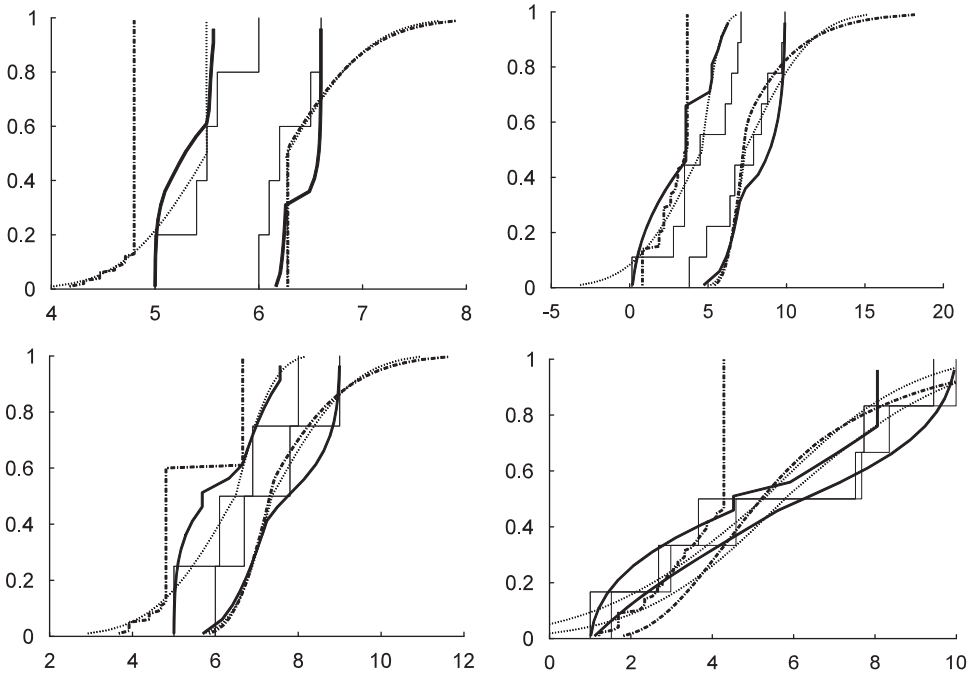

Fig. 7. Comparison of empirical, bounded Johnson, normal, and lognormal p-boxes for multiple interval examples (thick dashed line—lognormal, thin dotted line—normal, thick solid line—bounded Johnson, and thin solid line—empirical distribution). Top left, Example 1; top right, Example 2; bottom left, Example 3; bottom right, Example 4.

empirical p-box for all the four multiple interval examples, which is not necessarily the case for normal and lognormal p-boxes. One possible reason for this behavior could be the theoretical bounds that exist on normal, lognormal, and bounded Johnson distributions. The normal distribution is unbounded, and can lie between $[-\infty, +\infty]$, whereas the lognormal distribution is bounded between $[0, +\infty]$. The bounded Johnson distribution is restricted to lie within the interval bounds (discussed in Step 4 of Section 4.2).

As shown from the examples above, the proposed probabilistic representation of interval data using a family of bounded Johnson distributions is a viable approach for uncertainty quantification for interval uncertainty. Once such a family of distributions is constructed, it could be used in the context of uncertainty/reliability analysis using Monte Carlo simulations or FORM/SORM, resulting in set of values for an output quantity. This notion is unlike the case with aleatory uncertainties, where usually a single probabilistic representation describes the uncertainty, which yields a single quantity of interest from the uncertainty propagation stage. The proposed uncertainty representation is particularly suitable for use in FORM/SORM, since these methods require that the random variables are represented by probability distributions. These methods also require transforming the random variables into standard normal space, which is easy with Johnson distributions.

The state-of-the-art in uncertainty propagation in the presence of interval data requires a nested analysis – instances of interval variable are considered in an outer loop, each iteration of which requires a probabilistic analysis for the aleatory uncertainties – inner loop. Instead, one could use an *optimization-based* uncertainty propagation approach, where the parameters of the input interval variables (probabilistically described) that either maximize or minimize an output quantity of interest, e.g., probability of failure, can be found. We have proposed such optimization-based approaches for cases where the input variables are described by sparse point data [51]. Similar ideas can be extended to variables described by intervals, which will be studied in the future.

6. Concluding remarks

In this paper, we propose a probabilistic framework for representing uncertainty information available through interval data. The main contributions of this paper are: (1) development of algorithms to estimate bounds on the second, third, and fourth moments of single and multiple interval data, (2) demonstration that the proposed moment bounding algorithms are scalable in polynomial time, (3) use of the moment bounds thus estimated to fit a family of flexible Johnson distributions, (4) definition of a Johnson p-box, which is the bounding envelope of the family of Johnson distributions, and (5) development of an optimization-based method to construct the Johnson p-box.

Through scalability testing, we have shown that the algorithms to compute bounds on the second, third and fourth moment of interval data scale polynomially in the number of intervals. This is important because these problems have been generally considered earlier to be NP-hard. We have also shown how a probabilistic description for interval data can be provided by a *family of distributions.* Due to the nature of the interval data, however, we make no assumptions about the relative likelihood of any of these CDFs to be the true CDF. For point data, statistics such as moments or percentiles which are used to fit probability distributions assume single values. However, for interval data, we can only estimate bounds on the statistics such as moments or percentiles. Therefore, unlike for point data where there can be a single probability distribution as the uncertainty description, multiple probability distributions should describe interval data.

In this paper, we present an approach that can be used to fit a family of Johnson distributions using moment bounds obtained as discussed above. The family of Johnson distributions thus fit can be used as the probabilistic representation of the interval data. This process could also be performed using several other distributions. Johnson distributions offer an advantage because they have convenient transformations to be mapped into the normal space, which facilitates the use of popular analytical reliability methods such as FORM and SORM.

The proposed probabilistic framework of handling interval data can be applied for a combined treatment of aleatory and epistemic input uncertainties from the perspective of uncertainty propagation or reliability based design. This approach to uncertainty representation given interval data can allow for computationally efficient propagation by avoiding the nested analysis that is typically performed in the presence of interval variables.

Acknowledgement

This study was supported by funds from NASA Langley Research Center under Cooperative Agreement no. NNX08AF56A1 (Technical Monitor: Mr. Lawrence Green). The support is gratefully acknowledged.

Appendix 1

A.1. Johnson family of distributions

The Johnson family of distributions is parameterized by four parameters, γ, δ, λ, and ξ. Using these four parameters and a transformation function defined in Eq. (1) below, a standard normal distribution can be transformed into a Johnson distribution of four types—lognormal (S_L), unbounded Johnson (S_U), bounded Johnson (S_B), and normal distribution (S_N). Bounded Johnson distributions require the probability distribution to fall within a specified lower and upper bound, which may be a suitable distribution for certain physical quantities, such as Young's modulus. An unbounded Johnson distribution can theoretically lie between $-\infty$ and $+\infty$. Since the Johnson family of distributions has the flexibility to fit data with a large range of different distribution function shapes, this eliminates the need to test different distributions that will give the best fit to a set of sample data, especially when the data set is small.

Fitting Johnson distributions to sample data involves transforming a continuous random variable x, whose distribution is unknown, into a standard normal (z) with one of the four normalizing translations proposed by Johnson [52]. The general form of the translation is

$$z = \gamma + \delta f\left(\frac{x-\xi}{\lambda}\right) \tag{A1}$$

where $z \sim N(0,1)$, and f is the transformation corresponding to the Johnson distribution type (bounded, unbounded, normal, lognormal). The transformation functions that map different distributions to the standard normal distribution in the Johnson's family are as follows:

$$f(y) = \begin{cases} \ln(y) & \text{for lognormal}(S_L) \\ \ln\left(y+\sqrt{y^2+1}\right) & \text{for unbounded}(S_U) \\ \ln\left(\dfrac{y}{1-y}\right) & \text{for bounded}(S_B) \\ y & \text{for normal}(S_N) \end{cases} \tag{A2}$$

where $y=(x-\xi)/\lambda$. The problem of fitting Johnson distribution involves estimating the four parameters γ, δ, λ, and ξ, given the data set. Usually, various statistics of the data set, such as percentiles, quantiles, or moments are used to estímate the unknown parameters.

DeBrota et al. [53] present four methods to estimate the Johnson's parameters. The first method is the moment matching method. This method involves solving a set of four nonlinear equations that equate the first four moments calculated from the given data with those of a Johnson distribution, which are calculated as the function of unknown parameters. The second method is known as percentile matching, where the parameters are estimated by solving a system of nonlinear equations equating four percentile points of the data and the Johnson distribution. The third method involves a least squares estimation of the parameters obtained by minimizing the sum of squared errors in the percentile values from the data and those from the Johnson distribution. The fourth approach involves minimizing the error norm of the Johnson distribution CDF when compared with the empirical CDF constructed from the data. In this paper, we use the moment matching approach to fit Johnson distributions, in order to take advantage of the moment bounds calculated in Section 2.

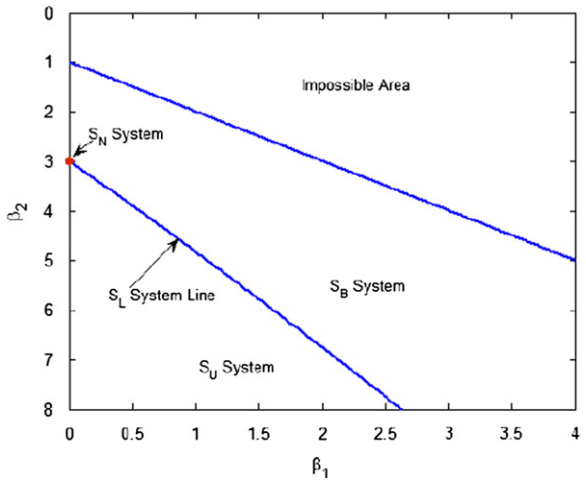

Fig. A1. Johnson distribution family identification.

The first step in fitting Johnson distributions is to determine what type (bounded, unbounded, lognormal, normal) is appropriate for the given data. The following standard procedure is available in the literature to find out the type of Johnson distribution [54] based on the moments of the data.

1. Calculate the second, third, fourth moments of the data: m_2, m_3, and m_4.
2. Calculate the ratios given below

$$\beta_1 \equiv m_3^2/m_2^3, \quad \beta_2 \equiv m_4/m_2^2$$

3. Use the chart in Fig. A1 to determine the appropriate distribution family.

We parenthetically note that for the bounded Johnson distribution, two of the unknown four parameters ξ and λ, are usually estimated from the bounds provided on the data. ξ is the lower bound on the data set and λ is the range of the data set.

References

[1] Parry GW. The characterization of uncertainty in probabilistic risk assessments of complex systems. Reliability Engineering and System Safety 1996;54:119–26.
[2] Baudrit C, Dubois D. Practical representations of incomplete probabilistic knowledge. Computational Statistics & Data Analysis 2006;51:86–108.
[3] Helton JC, Johnson JD, Oberkampf WL. An exploration of alternative approaches to the representation of uncertainty in model predictions. Reliability Engineering and System Safety 2004;85:39–71.
[4] Helton JC, Burmaster DE. Guest editorial: treatment of aleatory and epistemic uncertainty in performance assessments for complex systems. Reliability Engineering and System Safety 1996;54:91–4.
[5] Paté-Cornell ME. Uncertainties in risk analysis: six levels of treatment. Reliability Engineering and System Safety 1996;54(2–3):95–111.
[6] Ferson, S, Kreinovich, V, Hajagos, J, Oberkampf W, Ginzburg, L. Experimental uncertainty estimation and statistics for data having interval uncertainty.

Sandia National Laboratories Technical report SAND2007-0939, Albuquerque, NM, 2007.
[7] Du X, Sudjianto A, Huang B. Reliability based design with mixture of random and interval variables. Journal of Mechanical Design 2005;127:1068–76.
[8] Howson C, Urbach P. Scientific reasoning: the Bayesian approach. 2nd ed. Chicago, IL: Open Court; 1993.
[9] Bertrand P, Groupil F. Descriptive statistics for symbolic data. In: Bock HH, Diday E, editors. Analysis of symbolic data. Berlin: Springer; 2000.
[10] Williamson RC, Downs T. Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds. International Journal of Approximate Reasoning 1990;4:89–158.
[11] Regan HM, Ferson S, Berleant D. Equivalence of methods for uncertainty propagation of real-valued random variables. International Journal of Approximate Reasoning 2004;36:1–30.
[12] Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewicz SF, Ferson S. Challenge problems: uncertainty in system response given uncertain parameters. Reliability Engineering and System Safety 2004;85(1–3):11–9.
[13] Ferson S, Joslyn CA, Helton JC, Oberkampf WL, Sentz K. Summary from the epistemic uncertainty workshop: consensus amid diversity. Reliability Engineering and System Safety 2004;85:355–69.
[14] Klir GJ. Generalized information theory: aims, results, and open problems. Reliability Engineering and System Safety 2004;85:21–38.
[15] O'Hagan A, Oakley JE. Probability is perfect, but we can't elicit it perfectly. Reliability Engineering and System Safety 2004;85:239–48.
[16] Fetz T, Oberguggenberger M. Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. Reliability Engineering and System Safety 2004;85:73–87.
[17] Berleant D, Zhang J. Representation and problem solving with distribution envelope determination (DEnv). Reliability Engineering and System Safety 2004;85:153–68.
[18] Hall JW, Lawry J. Generation, combination and extension of random set approximations to coherent lower and upper probabilities. Reliability Engineering and System Safety 2004;85:89–101.
[19] Kozine IO, Utkin LV. An approach to combining unreliable pieces of evidence and their propagation in a system response analysis. Reliability Engineering and System Safety 2004;85:103–12.
[20] De Cooman G, Troffaes MCM. Coherent lower previsions in systems modeling: products and aggregation rules. Reliability Engineering and System Safety 2004;85:113–34.
[21] Ferson S, Hajagos J. Arithmetic with uncertain numbers: rigorous and (often) best-possible answers. Reliability Engineering and System Safety 2004;85:135–52.
[22] Red-Horse JR, Benjamin AS. A probabilistic approach to uncertainty quantification with limited information. Reliability Engineering and System Safety 2004;85:183–90.
[23] Ben-Haim Y. Uncertainty, probability and information-gaps. Reliability Engineering and System Safety 2004;85:249–66.
[24] Hailperin T. Boole's logic and probability. Amsterdam: North-Holland; 1986.
[25] Hyman, JM. FORSIG: an extension of FORTRAN with significance arithmetic. Report LA-9448-MS, Los Alamos National Laboratory, Los Alamos, NM, 1982. See also the website 〈http://math.lanl.gov/ams/report2000/significancearithmetic.html〉.
[26] Berleant D. Automatically verified reasoning with both intervals and probability density functions. Interval Computations 1993:48–70.
[27] Berleant D. Automatically verified arithmetic on probability distributions and intervals. In: Kearfott B, Kreinovich V, editors. Applications of interval computations. Kluwer Academic Publishers; 1996. p. 227–44.
[28] Berleant D, Goodman-Strauss C. Bounding the results of arithmetic operations on random variables of unknown dependency using intervals. Reliable Computing 1998;4:147–65.
[29] Helton,JC, Johnson JD, Oberkampf WL, Sallaberry CJ. Representation of analysis results involving aleatory and epistemic uncertainty, Sandia Report, SAND2008-4379, 2008.
[30] Apeland S, Aven T, Nilsen T. Quantifying uncertainty under a predictive, epistemic approach to risk analysis. Reliability Engineering and System Safety 2002;75:93–102.
[31] Hofer E, Kloos M, Hausmann BK, Peschke J, Woltereck M. An approximate epistemic uncertainty analysis approach in the presence of epistemic and aleatory uncertainties. Reliability Engineering and System Safety 2002;77:229–38.
[32] Agarwal H, Renaud JE, Preston EL, Padmanabhan D. Uncertainty quantification using evidence theory in multidisciplinary design optimization. Reliability Engineering and System Safety 2004;85(1–3):281–94.
[33] Shafer Glenn. A mathematical theory of evidence. Princeton University Press; 0-608-02508-9.
[34] Oberkampf, WL, Helton, JC, Sentz, K, 2001. Mathematical representation of uncertainty. In: Proceedings of the 42nd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference & exhibit, Seattle, WA, Paper number AIAA 2001-1645.
[35] Guo J, Du X. Sensitivity analysis with mixture of epistemic and aleatory variables. AIAA Journal 2007;45:9.
[36] Guo J, Du X. Reliability sensitivity analysis with random and interval variables. International Journal for Numerical Methods in Engineering 2009. [available online].
[37] Ben-Haim Y, Elishakoff I. Convex models of uncertainty in applied mechanics. Studies in Applied Mechanics 1990:25.
[38] Dubois D, Prade H. Possibility theory: an approach to computerized processing of uncertainty. 1st ed. Plenum Press; 1988.
[39] Rao SS, Annamdas KK. An evidence-based fuzzy approach for the safety analysis of uncertain systems. In: 50th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, Palm Springs, California, 2009, Paper number AIAA-2009-2263.
[40] Rutherford B. A response-modeling approach to characterization and propagation of uncertainty specified over intervals. Reliability Engineering and System Safety 2004;85:211–22.
[41] Helton JC, Johnson JD, Oberkampf WL, Storlie CB. A sampling-based computational strategy for the representation of epistemic uncertainty in model predictions with evidence theory. Comput. Methods Appl. Mech. Eng. 2007;196(37–40):3980–98.
[42] Penmetsa RC, Grandhi RV. Efficient estimation of reliability for problems with uncertain intervals. Computers and Structures 2002;80(12):1103–112.
[43] Rao SS, Cao L. Optimum design of mechanical systems involving interval parameters. Journal of Mechanical Design 2002;124(3):465–72.
[44] Langley RS. Unified approach to probabilistic and possibilistic analysis of uncertain systems. Journal of Engineering Mechanics 2000;126(11):1163–72.
[45] Du X, Chen W. An integrated methodology for uncertainty propagation and management in simulation-based systems design. AIAA Journal 2000;38(8):1471–8.
[46] Haldar A, Mahadevan S. Probability, reliability and statistical methods in engineering design. New York: John Wiley & Sons; 2000, 304pp., ISBN-10:0-471-33119-8.
[47] Kreinovich V. Probabilities, intervals, what next? Optimization problems related to extension of interval computations to situations with partial information about probabilities Journal of Global Optimization 2004;29(3):265–80.
[48] Kreinovich V, Xiang G, Starks SA, Longpre L, Ceberio M, Araiza R, et al. Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity. Reliable Computing 2006;12(4):273–80.
[49] DeGroot MH. Probability and statistics. 2nd ed. Addison-Wesley Publishing Company; 1984.
[50] DeBrota DJ, Roberts SD, Dittus RS, Wilson JR. Visual interactive fitting of bounded Johnson distributions. Simulation 1989;52(5):199–205.
[51] McDonald, M, Zaman, K, Mahadevan, S. Representation and first-order approximations for propagation of aleatory and distribution parameter uncertainty. In: Proceedings of the 50th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics, and materials conference, Palm Springs, California, 2009, Paper number AIAA-2009-2250.
[52] Johnson NL. Systems of frequency curves generated by methods of translation. Biometrika 1949;36:149–76.
[53] DeBrota, DJ, Swain, JJ, Roberts, SD, Venkataraman, S. Input modeling with the Johnson system of distributions. In: Proceedings of the 1988 winter simulation conference, 1988.
[54] Venkataraman S, Wilson JR. Modeling univariate populations with Johnson's translation system—description of the FITR1 software. Research Memorandum, School of Industrial Engineering, Purdue University, West Lafayette, IN, 1987.